Fabric Provisioning

This chapter contains these sections:

Fabric Provisioning

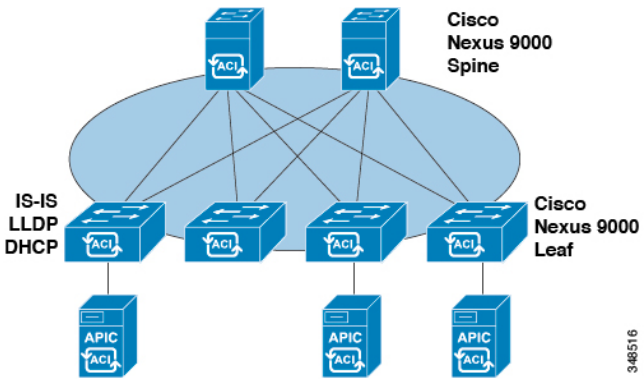
Cisco Application Centric Infrastructure (ACI) automation and self-provisioning offers these operation advantages over the traditional switching infrastructure:

- A clustered logically centralized but physically distributed APIC provides policy, bootstrap, and image management for the entire fabric.
- The APIC startup topology auto discovery, automated configuration, and infrastructure addressing uses these industry-standard protocols: Intermediate System-to-Intermediate System (IS-IS), Link Layer Discovery Protocol (LLDP), and Dynamic Host Configuration Protocol (DHCP).
- The APIC provides a simple and automated policy-based provisioning and upgrade process, and automated image management.
- APIC provides scalable configuration management. Because ACI data centers can be very large, configuring switches or interfaces individually does not scale well, even using scripts. APIC pod, controller, switch, module and interface selectors (all, range, specific instances) enable symmetric configurations across the fabric. To apply a symmetric configuration, an administrator defines switch profiles that associate interface configurations in a single policy group. The configuration is then rapidly deployed to all interfaces in that profile without the need to configure them individually.

Startup Discovery and Configuration

The clustered APIC controller provides DHCP, bootstrap configuration, and image management to the fabric for automated startup and upgrades. The following figure shows startup discovery.

Figure 1: Startup Discovery Configuration



The Cisco Nexus ACI fabric software is bundled as an ISO image, which can be installed on the Cisco APIC server through the KVM interface on the Cisco Integrated Management Controller (CIMC). The Cisco Nexus ACI Software ISO contains the Cisco APIC image, the firmware image for the leaf node, the firmware image for the spine node, default fabric infrastructure policies, and the protocols required for operation.

The ACI fabric bootstrap sequence begins when the fabric is booted with factory-installed images on all the switches. The Cisco Nexus 9000 Series switches that run the ACI firmware and APICs use a reserved overlay for the boot process. This infrastructure space is hard-coded on the switches. The APIC can connect to a leaf through the default overlay, or it can use a locally significant identifier.

The ACI fabric uses an infrastructure space, which is securely isolated in the fabric and is where all the topology discovery, fabric management, and infrastructure addressing is performed. ACI fabric management communication within the fabric takes place in the infrastructure space through internal private IP addresses. This addressing scheme allows the APIC to communicate with fabric nodes and other Cisco APIC controllers in the cluster. The APIC discovers the IP address and node information of other Cisco APIC controllers in the cluster using the Link Layer Discovery Protocol (LLDP)-based discovery process.

The following describes the APIC cluster discovery process:

- Each APIC in the Cisco ACI uses an internal private IP address to communicate with the ACI nodes and other APICs in the cluster. The APIC discovers the IP address of other APIC controllers in the cluster through the LLDP-based discovery process.
- APICs maintain an appliance vector (AV), which provides a mapping from an APIC ID to an APIC IP address and a universally unique identifier (UUID) of the APIC. Initially, each APIC starts with an AV filled with its local IP address, and all other APIC slots are marked as unknown.
- When a switch reboots, the policy element (PE) on the leaf gets its AV from the APIC. The switch then advertises this AV to all of its neighbors and reports any discrepancies between its local AV and neighbors' AVs to all the APICs in its local AV.

Using this process, the APIC learns about the other APIC controllers in the ACI through switches. After validating these newly discovered APIC controllers in the cluster, the APIC controllers update their local AV and program the switches with the new AV. Switches then start advertising this new AV. This process continues until all the switches have the identical AV and all APIC controllers know the IP address of all the other APIC controllers.

Note Prior to initiating a change to the cluster, always verify its health. When performing planned changes to the cluster, all controllers in the cluster should be healthy. If one or more of the APIC controllers in the cluster is not healthy, remedy that situation before proceeding with making changes to the cluster. Also, assure that cluster controllers added to the APIC are running the same version of firmware as the other controllers in the APIC cluster. See the KB: Cisco ACI APIC Cluster Management article for guidelines that must be followed to assure that making changes the APIC cluster complete normally.

The ACI fabric is brought up in a cascading manner, starting with the leaf nodes that are directly attached to the APIC. LLDP and control-plane IS-IS convergence occurs in parallel to this boot process. The ACI fabric uses LLDP- and DHCP-based fabric discovery to automatically discover the fabric switch nodes, assign the infrastructure VXLAN tunnel endpoint (VTEP) addresses, and install the firmware on the switches. Prior to this automated process, a minimal bootstrap configuration must be performed on the Cisco APIC controller. After the APIC controllers are connected and their IP addresses assigned, the APIC GUI can be accessed by entering the address of any APIC controller into a web browser. The APIC GUI runs HTML5 and eliminates the need for Java to be installed locally.

Fabric Inventory

The policy model contains a complete real-time inventory of the fabric, including all nodes and interfaces. This inventory capability enables automation of provisioning, troubleshooting, auditing, and monitoring.

For Cisco ACI fabric switches, the fabric membership node inventory contains policies that identify the node ID, serial number, and name. Third-party nodes are recorded as unmanaged fabric nodes. Cisco ACI switches can be automatically discovered, or their policy information can be imported. The policy model also maintains fabric member node state information.

Node States	Condition
Unknown	No policy. All nodes require a policy; without a policy, a member node state is unknown.
Discovering	A transient state showing that the node is being discovered and waiting for host traffic.
Undiscovered	The node has policy but has never been brought up in the fabric.
Unsupported	The node is a Cisco switch but it is not supported. For example, the firmware version is not compatible with ACI fabric.
Decommissioned	The node has a policy, was discovered, but a user disabled it. The node can be reenabled. **Note** Specifying the wipe option when decommissioning a leaf switch results in the APIC attempting to remove all the leaf switch configurations on both the leaf switch and on the APIC. If the leaf switch is not reachable, only the APIC is cleaned. In this case, the user must manually wipe the leaf switch by resetting it.
Inactive	The node is unreachable. It had been discovered but currently is not accessible. For example, it may be powered off, or its cables may be disconnected.
Active	The node is an active member of the fabric.

Disabled interfaces can be ones blacklisted by an administrator or ones taken down because the APIC detects anomalies. Examples of link state anomalies include the following:

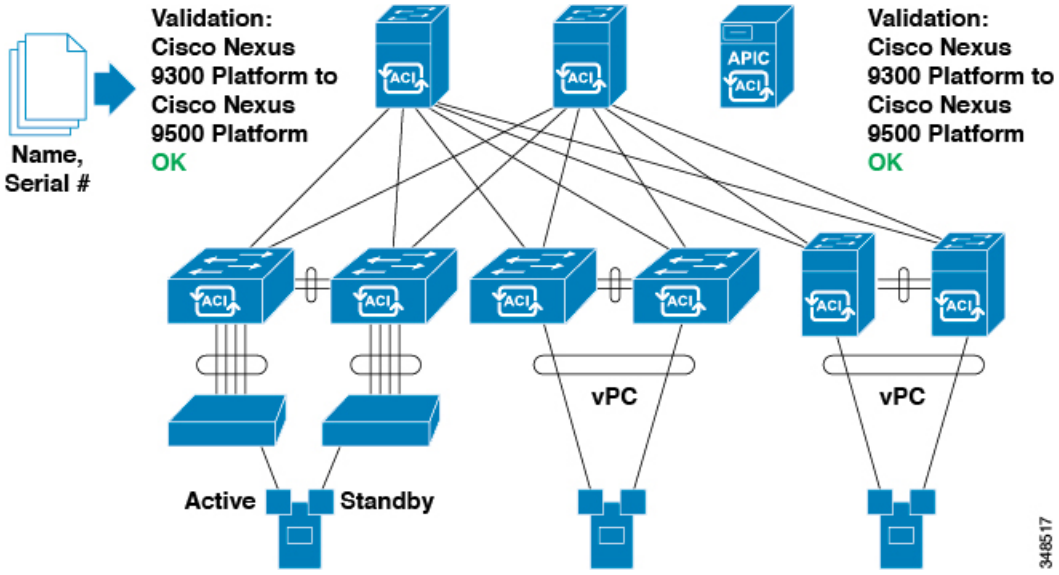
- A wiring mismatch, such as a spine connected to a spine, a leaf connected to a leaf, a spine connected to a leaf access port, a spine connected to a non-ACI node, or a leaf fabric port connected to a non-ACI device.
- A fabric name mismatch. The fabric name is stored in each ACI node. If a node is moved to another fabric without resetting it to a back to factory default state, it will retain the fabric name.
- A UUID mismatch causes the APIC to disable the node.

Note If an administrator uses the APIC to disable all the leaf nodes on a spine, a spine reboot is required to recover access to the spine.

Provisioning

The APIC provisioning method automatically brings up the ACI fabric with the appropriate connections. The following figure shows fabric provisioning.

Figure 2: Fabric Provisioning



After Link Layer Discovery Protocol (LLDP) discovery learns all neighboring connections dynamically, these connections are validated against a loose specification rule such as "LEAF can connect to only SPINE-L1-*" or "SPINE-L1-* can connect to SPINE-L2-* or LEAF." If a rule mismatch occurs, a fault occurs and the connection is blocked because a leaf is not allowed to be connected to another leaf, or a spine connected to a spine. In addition, an alarm is created to indicate that the connection needs attention. The Cisco ACI fabric administrator can import the names and serial numbers of all the fabric nodes from a text file into the APIC or allow the fabric to discover the serial numbers automatically and then assign names to the nodes using the APIC GUI, command-line interface (CLI), or API. The APIC is discoverable via SNMP. It has the following asysobjectId: `ciscoACIController OBJECT IDENTIFIER ::= { ciscoProducts 2238 }`

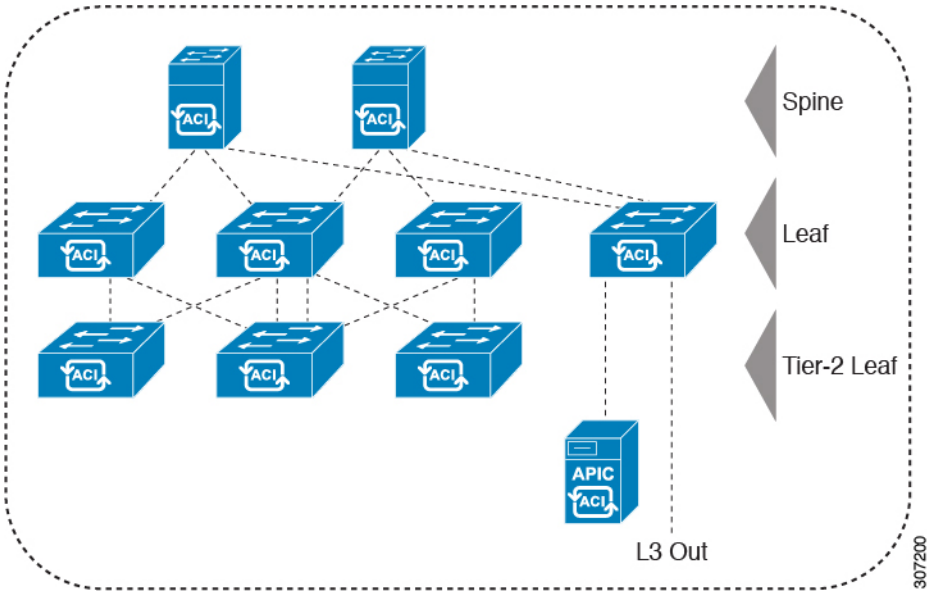
Multi-Tier Architecture

3-tier Core-Aggregation-Access architectures are common in data center network topologies. As of the Cisco APIC Release 4.1(1), you can create a multi-tier ACI fabric topology that corresponds to the Core-Aggregation-Access architecture, thus mitigating the need to upgrade costly components such as rack space or cabling. The addition of a tier-2 leaf layer makes this topology possible. The tier-2 leaf layer supports connectivity to hosts or servers on the downlink ports and connectivity to the leaf layer (aggregation) on the uplink ports.

In the multi-tier topology, the leaf switches initially have uplink connectivity to the spine switches and downlink connectivity to the tier-2 leaf switches. To make the entire topology an ACI fabric, all ports on the leaf switches connecting to tier-2 leaf fabric ports must be configured as fabric ports (if not already using the default fabric ports). After APIC discovers the tier-2 leaf switch, you can change the downlink port on the tier-2 leaf to a fabric port and connect to an uplink port on the middle layer leaf.

The following figure shows an example of a multi-tier fabric topology.

Figure 3: Multi-Tier Fabric Topology Example



While the topology in the above image shows the Cisco APIC and L3Out/EPG connected to the leaf aggregation layer, the tier-2 leaf access layer also supports connectivity to APICs and L3Out/EPGs.

APIC Cluster Management

Best practices for cluster management

- Always verify the health of all controllers before making any changes to the cluster. Confirm that every controller is fully fit and resolve any health issues before proceeding.
- Ensure that all controllers in the cluster run the same firmware version before adding, configuring, or clustering devices. Do not cluster controllers running different firmware versions.
- Maintain at least three active controllers in your cluster, and add standby controllers as needed. For scalability requirements, consult the Verified Scalability Guide to determine the required number of active controllers for your deployment.
- Ignore cluster information from controllers that are not currently active in the cluster, as their data may be inaccurate.
- Know that once you configure a cluster slot with a controller's ChassisID, you must decommission that controller to make the slot available for reassignment.
- Wait for all ongoing firmware upgrades to complete and verify the cluster is fully fit before making additional changes.
- When moving a controller, always ensure the cluster is healthy. Select the controller you intend to move, shut it down, physically move and reconnect it, and then power it on. After the move, verify through the management interface that all controllers return to a fully fit state.

- Move only one controller at a time to maintain cluster stability.
- When transferring a controller to a different set of leaf switches or to a different port within the same leaf switch, ensure the cluster is healthy first. Decommission the controller before moving it, and then recommission it after the move.
- Before configuring the cluster, confirm that all controllers run the same firmware version to prevent unsupported operations and cluster issues.
- Delete any unused OOB EPGs associated with a controller. Assigning multiple EPGs to a controller is not supported and can cause the cluster workflow IP address to be overridden by policy.
- Remember that log record objects are stored only in one shard on a single controller. If you decommission or replace that controller, those logs are permanently lost.
- When decommissioning a controller, be aware that all fault, event, and audit log history stored on it is deleted. If you replace all controllers, all log history is lost. Before migrating a controller, manually back up its log history to prevent data loss.

Cold standby clusters

A cold standby cluster is a high availability deployment that

- operates with some controllers as active and others as standby,
- enables standby controllers to quickly take over when an active controller fails, and
- allows administrators to manually initiate a switchover to maintain cluster services.

Additional information

In a Cisco APIC cluster, administrators typically configure at least three active controllers and one or more standby controllers for resilience. The active controllers manage traffic under normal conditions, while standby controllers remain ready to replace any active controller if needed. The switch to a standby controller is performed manually to ensure controlled recovery during failures. This arrangement maintains network stability and reduces downtime.

Maintenance Mode

Following are terms that are helpful to understand when using maintenance mode:

- **Maintenance mode**: Used to isolate a switch from user traffic for debugging purposes. You can put a switch in **maintenance mode** by enabling the **Maintenance (GIR)** field in the **Fabric Membership** page in the APIC GUI, located at **Fabric** > **Inventory** > **Fabric Membership** (right-click on a switch and choose **Maintenance (GIR)**).

  If you put a switch in **maintenance mode**, that switch is not considered as a part of the operational ACI fabric infra and it will not accept regular APIC communications.

You can use maintenance mode to gracefully remove a switch and isolate it from the network in order to perform debugging operations. The switch is removed from the regular forwarding path with minimal traffic disruption.

In graceful removal, all external protocols are gracefully brought down except the fabric protocol (IS-IS) and the switch is isolated from the network. During maintenance mode, the maximum metric is advertised in IS-IS within the Cisco Application Centric Infrastructure (Cisco ACI) fabric and therefore the leaf switch in maintenance mode does not attract traffic from the spine switches. In addition, all front-panel interfaces on the switch are shutdown except for the fabric interfaces. To return the switch to its fully operational (normal) mode after the debugging operations, you must recommission the switch. This operation will trigger a stateless reload of the switch.

In graceful insertion, the switch is automatically decommissioned, rebooted, and recommissioned. When recommissioning is completed, all external protocols are restored and maximum metric in IS-IS is reset after 10 minutes.

The following protocols are supported:

- Border Gateway Protocol (BGP)
- Enhanced Interior Gateway Routing Protocol (EIGRP)
- Intermediate System-to-Intermediate System (IS-IS)
- Open Shortest Path First (OSPF)
- Link Aggregation Control Protocol (LACP)

Protocol Independent Multicast (PIM) is not supported.

Important Notes

- If a border leaf switch has a static route and is placed in maintenance mode, the route from the border leaf switch might not be removed from the routing table of switches in the ACI fabric, which causes routing issues.

  To work around this issue, either:

  - Configure the same static route with the same administrative distance on the other border leaf switch, or
  - Use IP SLA or BFD for track reachability to the next hop of the static route

- While the switch is in maintenance mode, the Ethernet port module stops propagating the interface related notifications. As a result, if the remote switch is rebooted or the fabric link is flapped during this time, the fabric link will not come up afterward unless the switch is manually rebooted (using the **acidiag touch clean** command), decommissioned, and recommissioned.
- While the switch is in maintenance mode, CLI 'show' commands on the switch show the front panel ports as being in the up state and the BGP protocol as up and running. The interfaces are actually shut and all other adjacencies for BGP are brought down, but the displayed active states allow for debugging.
- For multi-pod / multi-site, **IS-IS metric for redistributed routes** should be set to less than 63 to minimize the traffic disruption when bringing the node back into the fabric. To set the **IS-IS metric for redistributed routes**, choose **Fabric** > **Fabric Policies** > **Pod Policies** > **IS-IS Policy**.
- When you reboot a spine or leaf and after the IS-IS adjacency comes up the **IS-IS metric for redistributed routes** is advertised as high, which is, 34 and will not be available as an ECMP next hop.
- Existing GIR supports all Layer 3 traffic diversion. With LACP, all the Layer 2 traffic is also diverted to the redundant node. Once a node goes into maintenance mode, LACP running on the node immediately

informs neighbors that it can no longer be aggregated as part of port-channel. All traffic is then diverted to the vPC peer node.

- The following operations are not allowed in maintenance mode:
  - **Upgrade**: Upgrading the network to a newer version
  - **Stateful Reload**: Restarting the GIR node or its connected peers
  - **Stateless Reload**: Restarting with a clean configuration or power-cycle of the GIR node or its connected peers
  - **Link Operations**: Shut / no-shut or optics OIR on the GIR node or its peer node
  - **Configuration Change**: Any configuration change (such as clean configuration, import, or snapshot rollback)
  - **Hardware Change**: Any hardware change (such as adding, swapping, removing FRU's or RMA)

Leaf reload in the absence of Cisco APIC

An ungraceful reload occurs when Cisco ACI switches restart unexpectedly, such as during a power loss or crash, instead of a user-initiated reload. In releases prior to Cisco ACI 6.2(1), switches that perform an ungraceful reload start in a stateless state. They attempt to be discovered by the APICs and download their configurations. If the APICs are down or unreachable, the switches cannot restore their configuration and will remain nonfunctional until the APICs are accessible.

Starting with Cisco ACI release 6.2(1), Cisco ACI switches retain a snapshot of its configuration which will be persistent across ungraceful reload. If an ungraceful reload occurs, the switch restores its configuration from the most recent snapshot without having to download it from APIC. If there was any configuration changes that occurred on APIC while the switch was down or between the snapshot schedule, the switch reconciles its configuration with the latest version on the APICs. This new mechanism prevents the switch from getting stuck in an unoperational state indefinitely while APICs are down or not reachable for any reason.

Cisco APIC triggers a snapshot of the switch configuration during certain events, such as when the switch joins or rejoins the fabric or when configuration changes are made on the APICs.

Collecting snapshots

When a switch joins the APIC cluster for the first time, the APIC initiates a baseline snapshot collection on the switch. The initial snapshot is taken 15 minutes after the node joins the cluster. Additional snapshots are captured every 15 minutes until there are ten in total. Any configuration changes in APIC will trigger up to two snapshots within each 24-hour period, after the initial set is complete.

Figure 4: Baseline snapshot collection time line

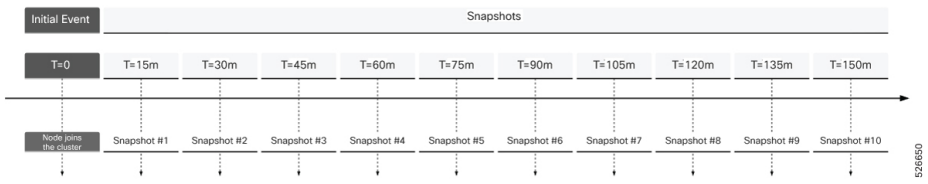
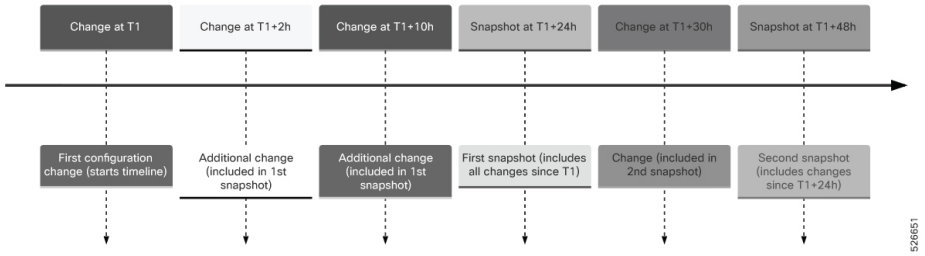


Figure 5: Configuration change snapshot collection time line



Example of a baseline snapshot

The example below shows the Switch CLI output from the first snapshot.

```
ins-aci-vapic1# show switch snapshot detail
 Node ID  Node Name             Serial          Role    Total Attempts  Completed Count
Failure Count  Failure Reason   State           Last Attempt Timestamp         Last Success
 Timestamp
 ----     --------------------  --------------  ------  ----            ----
----           ---------------  ---------------  ------------------------------
------------------------------
 101      ins-vc51-leaf1        FLM26480CGJ     leaf    10              10
10            NoFailure        completed       2025-08-28T06:09:50.510+00:00
2025-08-28T06:09:50.510+00:00
 201      ins-vc51-spine1       FDO27151XQN     spine   10              10
10            NoFailure        completed       2025-08-28T06:18:51.374+00:00
2025-08-28T06:18:51.374+00:0
```

Example of a snapshot taken 24 Hours after node configuration change

The example below shows the Switch CLI output taken 24 hours after the node's configuration change.

```
ins-aci-vapic1# show switch snapshot detail
 Node ID  Node Name             Serial          Role    Total Attempts  Completed Count
Failure Count  Failure Reason   State           Last Attempt Timestamp         Last Success
 Timestamp
 ----     --------------------  --------------  ------  ----            ----
----           ---------------  ---------------  ------------------------------
------------------------------
 101      ins-vc51-leaf1        FLM26480CGJ     leaf    2               2                2
          NoFailure        completed       2025-08-29T07:09:50.510+00:00
2025-08-29T07:09:50.510+00:00
 201      ins-vc51-spine1       FDO27151XQN     spine   2               2                2
          NoFailure        completed       2025-08-29T07:18:51.374+00:00
2025-08-29T07:18:51.374+00:00
```

Handling different reload scenarios

Stateless reload

A stateless reload works like a factory reset. All existing configurations on the switch are erased. After the switch reboots, it must download all policies and configurations from the controller before it can operate. The switch starts without any saved state and retrieves the necessary policies from the APIC to resume functioning. Any snapshots captured on the switch are deleted during this process.

To perform a stateless reload, use the following commands:

```
leaf1# setup-clean-config.sh

leaf1# reload
```

Stateful reload

When you run the reload command, the switch restarts but keeps its current configuration. After rebooting, the switch can resume forwarding traffic right away without needing to download configurations from the controller. If the reload is ungraceful, the switch tries to restore its state from the latest available snapshot.

To perform a stateful reload, use the following command:

```
leaf1# reload
```

Ungraceful reload

An ungraceful reload is an unintended reload due to power loss or crash. The new handling of ungraceful reload with the internal snapshot capability was introduced to handle the ungraceful reload as if it were a stateful reload as explained above.

Limitations

This limitation applies when a leaf switch reloads and Cisco APIC is unavailable.

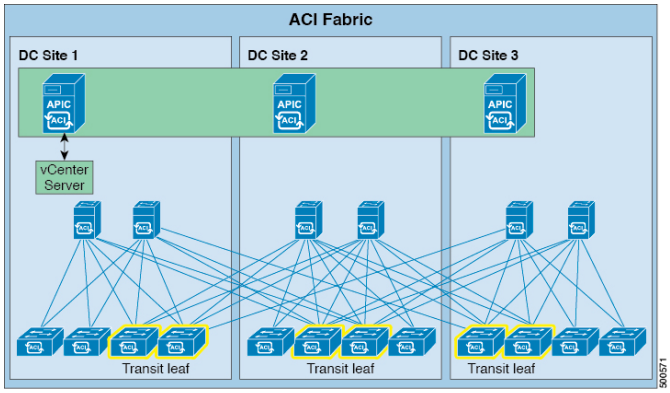
- When the port is brought up, a snapshot is taken every 24 hours. If you make a configuration change—such as adding a static vPC—that is not included in the current snapshot, the downlink port may come up after a power cycle but will experience traffic loss.

Stretched ACI Fabric Design Overview

Stretched ACI fabric is a partially meshed design that connects ACI leaf and spine switches distributed in multiple locations. Typically, an ACI fabric implementation is a single site where the full mesh design connects each leaf switch to each spine switch in the fabric, which yields the best throughput and convergence. In multi-site scenarios, full mesh connectivity may be not possible or may be too costly. Multiple sites, buildings, or rooms can span distances that are not serviceable by enough fiber connections or are too costly to connect each leaf switch to each spine switch across the sites.

The following figure illustrates a stretched fabric topology.

Figure 6: ACI Stretched Fabric Topology



The stretched fabric is a single ACI fabric. The sites are one administration domain and one availability zone. Administrators are able to manage the sites as one entity; configuration changes made on any APIC controller node are applied to devices across the sites. The stretched ACI fabric preserves live VM migration capability

across the sites. The ACI stretched fabric design has been validated, and is hence supported, on up to three interconnected sites.

An ACI stretched fabric essentially represents a "stretched pod" extended across different locations. A more solid, resilient (and hence recommended) way to deploy an ACI fabric in a distributed fashion across different locations is offered since ACI release 2.0(1) with the ACI Multi-Pod architecture. For more information, refer to the following white paper:

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

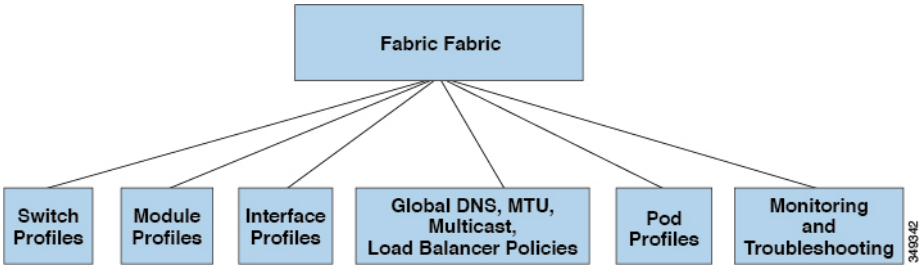
Stretched ACI Fabric Related Documents

The KB Stretched ACI Fabric Design Overview technical note provides design guidelines regarding traffic flow, APIC cluster redundancy and operational considerations for implementing an ACI fabric stretched across multiple sites.

Fabric Policies Overview

Fabric policies govern the operation of internal fabric interfaces and enable the configuration of various functions, protocols, and interfaces that connect spine and leaf switches. Administrators who have fabric administrator privileges can create new fabric policies according to their requirements. The APIC enables administrators to select the pods, switches, and interfaces to which they will apply fabric policies. The following figure provides an overview of the fabric policy model.

Figure 7: Fabric Polices Overview



Fabric policies are grouped into the following categories:

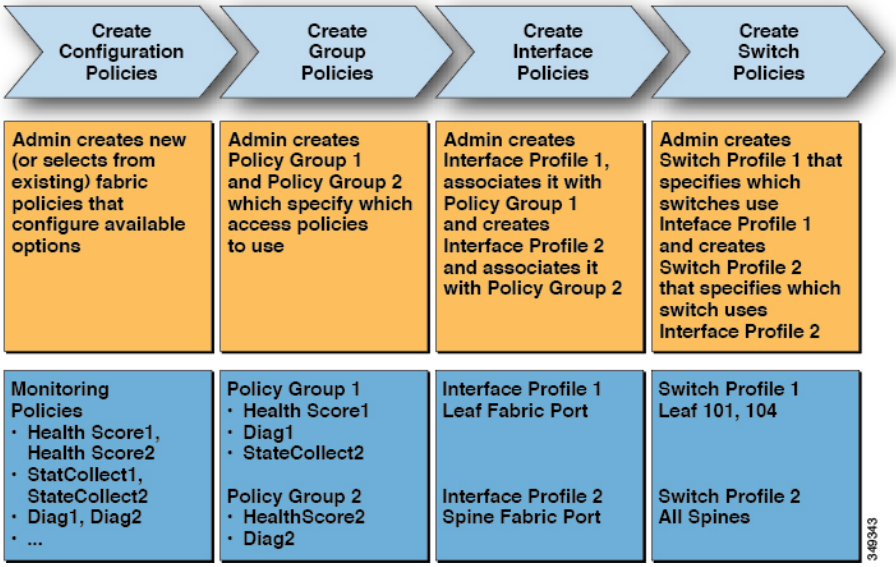
- Switch profiles specify which switches to configure and the switch configuration policy.
- Module profiles specify which spine switch modules to configure and the spine switch configuration policy.
- Interface profiles specify which fabric interfaces to configure and the interface configuration policy.
- Global policies specify DNS, fabric MTU default, multicast tree, and load balancer configurations to be used throughout the fabric.
- Pod profiles specify date and time, SNMP, council of oracle protocol (COOP), IS-IS and Border Gateway Protocol (BGP) route reflector policies.
- Monitoring and troubleshooting policies specify what to monitor, thresholds, how to handle faults and logs, and how to perform diagnostics.

Fabric Policy Configuration

Fabric policies configure interfaces that connect spine and leaf switches. Fabric policies can enable features such as monitoring (statistics collection and statistics export), troubleshooting (on-demand diagnostics and SPAN), IS-IS, council of oracle protocol (COOP), SNMP, Border Gateway Protocol (BGP) route reflectors, DNS, or Network Time Protocol (NTP).

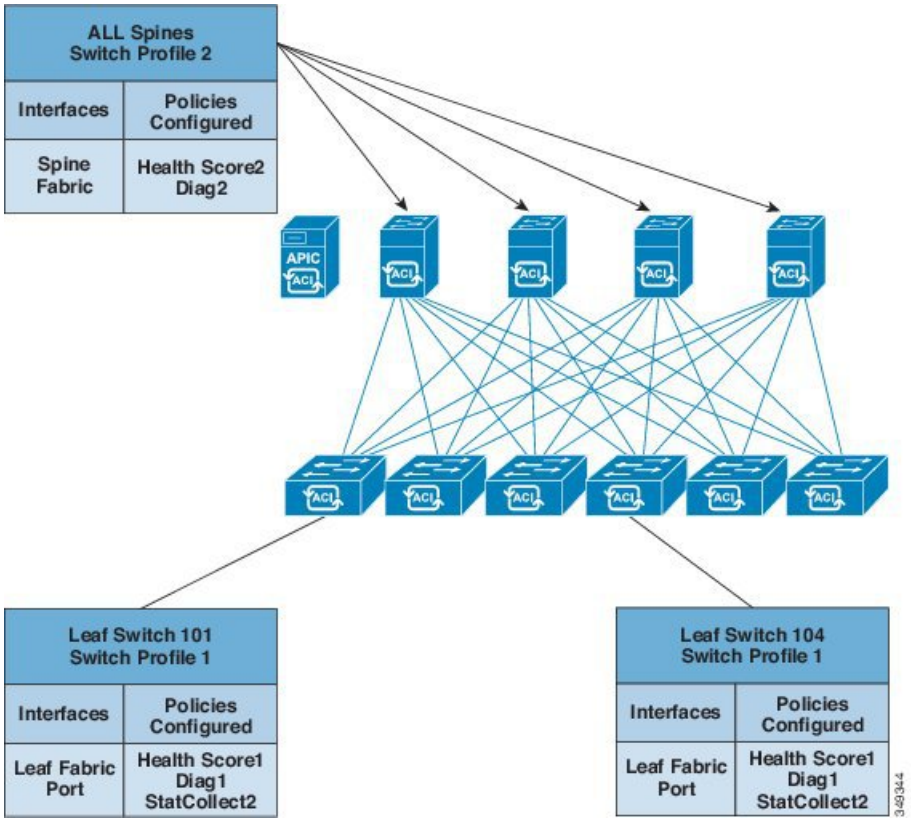
To apply a configuration across the fabric, an administrator associates a defined group of policies to interfaces on switches in a single step. In this way, large numbers of interfaces across the fabric can be configured at once; configuring one port at a time is not scalable. The following figure shows how the process works for configuring the ACI fabric.

Figure 8: Fabric Policy Configuration Process



The following figure shows the result of applying Switch Profile 1 and Switch Profile 2 to the ACI fabric.

Figure 9: Application of a Fabric Switch Policy



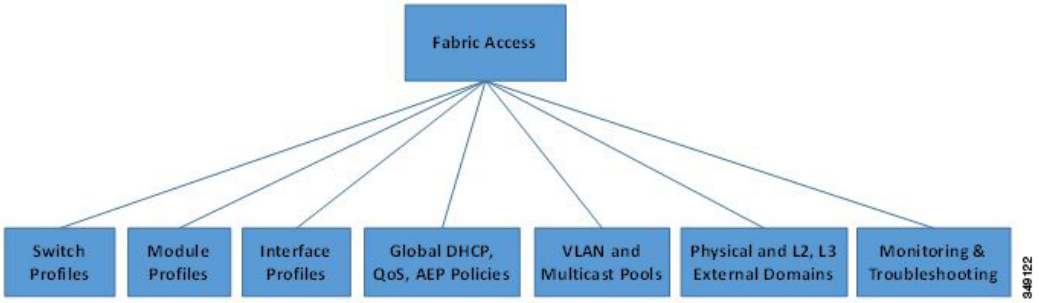
This combination of infrastructure and scope enables administrators to manage fabric configuration in a scalable fashion. These configurations can be implemented using the REST API, the CLI, or the GUI. The Quick Start Fabric Interface Configuration wizard in the GUI automatically creates the necessary underlying objects to implement such policies.

Access Policies Overview

Access policies configure external-facing interfaces that connect to devices such as virtual machine controllers and hypervisors, hosts, network attached storage, routers, or Fabric Extender (FEX) interfaces. Access policies enable the configuration of port channels and virtual port channels, protocols such as Link Layer Discovery Protocol (LLDP), Cisco Discovery Protocol (CDP), or Link Aggregation Control Protocol (LACP), and features such as statistics gathering, monitoring, and diagnostics.

The following figure provides an overview of the access policy model.

Figure 10: Access Policy Model Overview



Access policies are grouped into the following categories:

- Switch profiles specify which switches to configure and the switch configuration policy.
- Module profiles specify which leaf switch access cards and access modules to configure and the leaf switch configuration policy.
- Interface profiles specify which access interfaces to configure and the interface configuration policy.
- Global policies enable the configuration of DHCP, QoS, and attachable access entity (AEP) profile functions that can be used throughout the fabric. AEP profiles provide a template to deploy hypervisor policies on a large set of leaf ports and associate a Virtual Machine Management (VMM) domain and the physical network infrastructure. They are also required for Layer 2 and Layer 3 external network connectivity.
- Pools specify VLAN, VXLAN, and multicast address pools. A pool is a shared resource that can be consumed by multiple domains such as VMM and Layer 4 to Layer 7 services. A pool represents a range of traffic encapsulation ide.jpgiers (for example, VLAN IDs, VNIDs, and multicast addresses).
- Physical and external domains policies include the following:
  - External bridged domain Layer 2 domain profiles contain the port and VLAN specifications that a bridged Layer 2 network connected to the fabric uses.
  - External routed domain Layer 3 domain profiles contain the port and VLAN specifications that a routed Layer 3 network connected to the fabric uses.
  - Physical domain policies contain physical infrastructure specifications, such as ports and VLAN, used by a tenant or endpoint group.
- Monitoring and troubleshooting policies specify what to monitor, thresholds, how to handle faults and logs, and how to perform diagnostics.

Access Policy Configuration

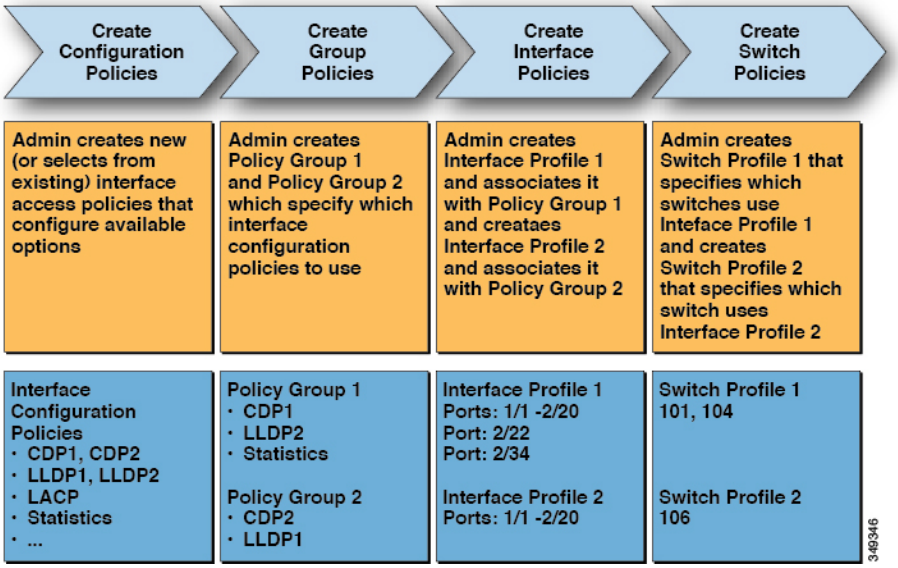
Access policies configure external-facing interfaces that do not connect to a spine switch. External-facing interfaces connect to external devices such as virtual machine controllers and hypervisors, hosts, routers, or Fabric Extenders (FEXs). Access policies enable an administrator to configure port channels and virtual port channels, protocols such as LLDP, CDP, or LACP, and features such as monitoring or diagnostics.

Sample XML policies for switch interfaces, port channels, virtual port channels, and change interface speeds are provided in *Cisco APIC Rest API Configuration Guide*.

Note While tenant network policies are configured separately from fabric access policies, tenant policies are not activated unless the underlying access policies they depend on are in place.

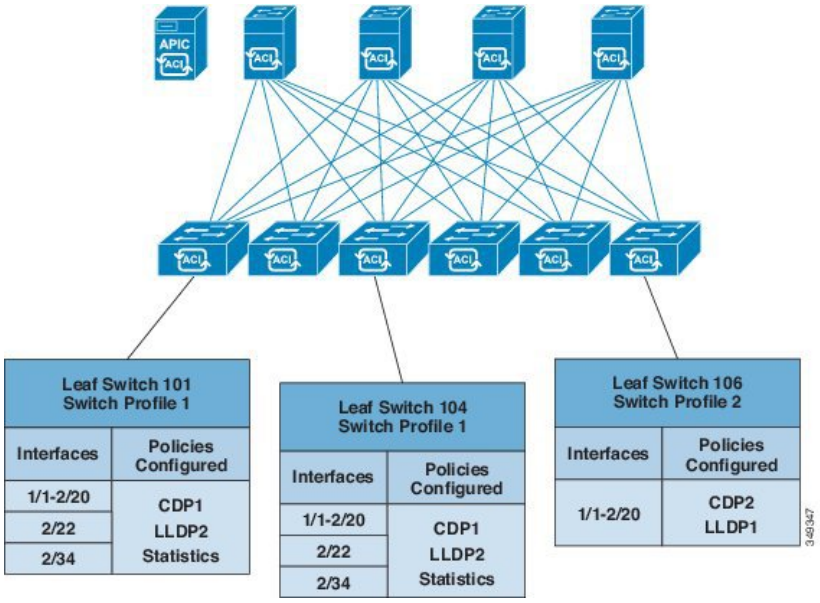
To apply a configuration across a potentially large number of switches, an administrator defines switch profiles that associate interface configurations in a single policy group. In this way, large numbers of interfaces across the fabric can be configured at once. Switch profiles can contain symmetric configurations for multiple switches or unique special purpose configurations. The following figure shows the process for configuring access to the ACI fabric.

Figure 11: Access Policy Configuration Process



The following figure shows the result of applying Switch Profile 1 and Switch Profile 2 to the ACI fabric.

Figure 12: Applying an Access Switch Policy

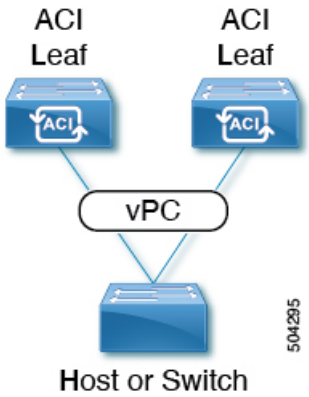


This combination of infrastructure and scope enables administrators to manage fabric configuration in a scalable fashion. These configurations can be implemented using the REST API, the CLI, or the GUI. The Quick Start Interface, PC, VPC Configuration wizard in the GUI automatically creates the necessary underlying objects to implement such policies.

Virtual Port Channels in Cisco ACI

A virtual port channel (vPC) allows links that are physically connected to two different Cisco Application Centric Infrastructure (ACI) leaf nodes to appear as a single port channel (PC) to a third device, such as a network switch, server, any other networking device that supports link aggregation technology. vPCs consist of two Cisco ACI leaf switches designated as vPC peer switches. Of the vPC peers, one is primary and one is secondary. The system formed by the switches is referred to as a vPC domain.

Figure 13: vPC Domain



The following behavior is specific to the Cisco ACI vPC implementation:

- No dedicated peer-link between the vPC peers. Instead, the fabric itself serves as the Multi-Chassis Trunking (MCT).
- Peer reachability protocol: Cisco ACI uses the Zero Message Queue (ZMQ) instead of Cisco Fabric Services (CFS).
  - ZMQ is an open-source, high-performance messaging library that uses TCP as the transport.
  - This library is packaged as libzmq on the switch and linked into each application that needs to communicate with a vPC peer.
- Peer reachability is not handled using a physical peer link. Instead, routing triggers are used to detect peer reachability.
  - The vPC manager registers with Unicast Routing Information Base (URIB) for peer route notifications.
  - When IS-IS discovers a route to the peer, URIB notifies the vPC manager, which in turn attempts to open a ZMQ socket with the peer.
  - When the peer route is withdrawn by IS-IS, the vPC manager is again notified by URIB, and the vPC manager brings down the MCT link.
- When creating a vPC domain between two leaf switches, the following hardware model limitations apply:
  - Generation 1 switches are compatible only with other generation 1 switches. These switch models can be identified by the lack of "EX," "FX, "FX2," "GX," or later suffix at the end of the switch name. For example, N9K-9312TX.
  - Generation 2 and later switches can be mixed together in a vPC domain. These switch models can be identified by the "EX," "FX, "FX2," "GX," or later suffix at the end of the switch name. For example N9K-93108TC-EX or N9K-9348GC-FXP.

  Examples of compatible vPC switch pairs:
  - N9K-C9312TX and N9K-C9312TX
  - N9K-C93108TC-EX and N9K-C9348GC-FXP
  - N9K-C93180TC-FX and N9K-C93180YC-FX
  - N9K-C93180YC-FX and N9K-C93180YC-FX

  Examples of incompatible vPC switch pairs:
  - N9K-C9312TX and N9K-C93108TC-EX
  - N9K-C9312TX and N9K-C93180YC-FX
- Port channels and virtual port channels can be configured with or without LACP.

  If you configure a virtual port channel with LACP, LACP sets a port to the suspended state if it does not receive an LACP PDU from the peer. This can cause some servers to fail to boot up as they require LACP to bring up the port logically. You can tune the behavior to individual use by disabling **LACP suspend individual**. To do so, create a port channel policy in your vPC policy group, and after setting the mode to LACP active, remove **Suspend Individual Port**. Afterward, the ports in the vPC will stay active and continue to send LACP packets.

- Adaptive load balancing (ALB), based on ARP negotiation, across virtual port channels is not supported in Cisco ACI.
- The 25 gig port does not come up on its own and displays the **vpc peerlink is down** error, if one of the vPC links is not cabled correctly.

Port Channel and Virtual Port Channel Access

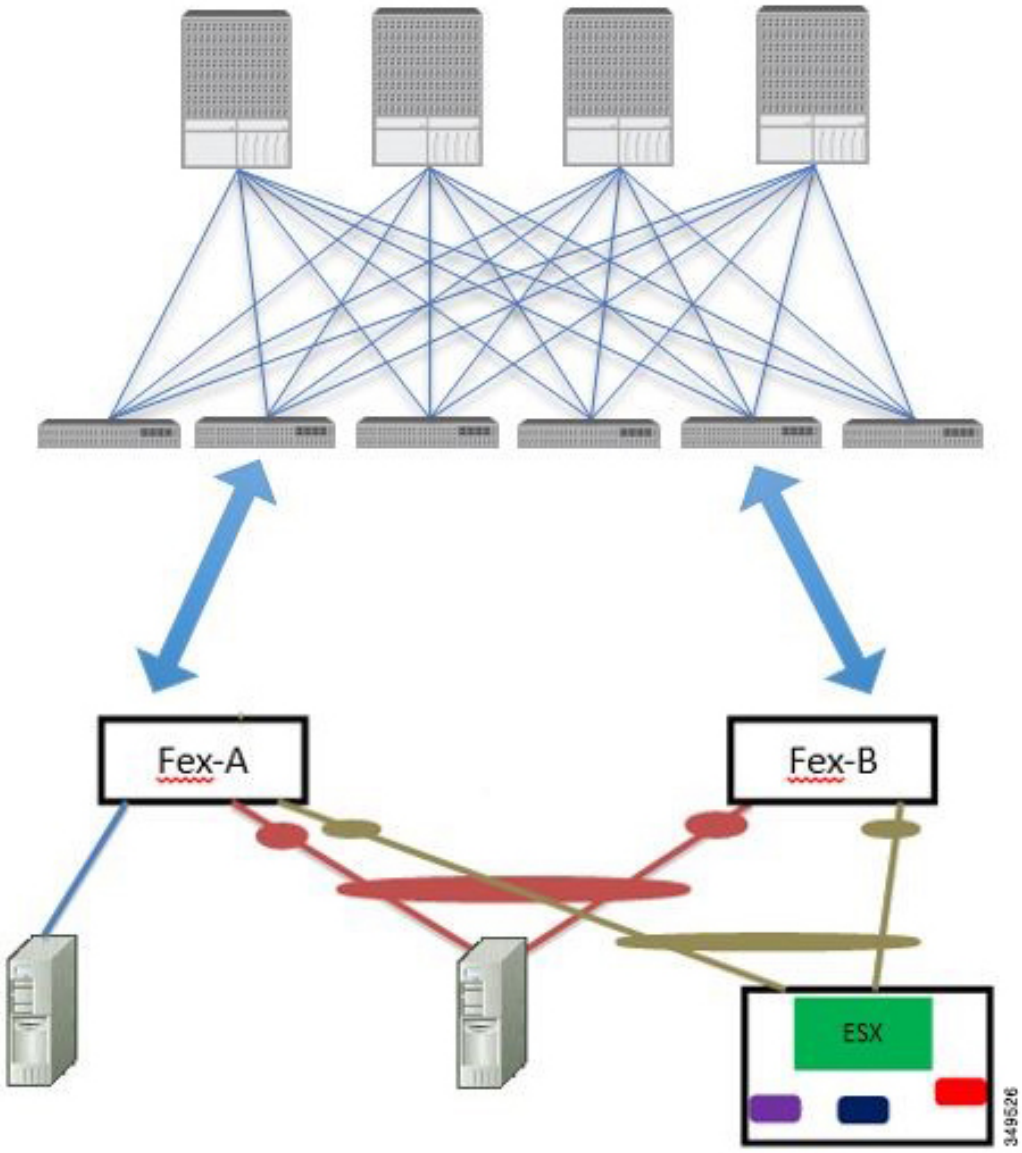
Access policies enable an administrator to configure port channels and virtual port channels. Sample XML policies for switch interfaces, port channels, virtual port channels, and change interface speeds are provided in *Cisco APIC Rest API Configuration Guide*.

FEX Virtual Port Channels

The ACI fabric supports Cisco Fabric Extender (FEX) server-side virtual port channels (vPC), also known as an FEX straight-through vPC.

Note When creating a vPC domain between two leaf switches, ensure that the hardware of the two leaf switches that are going to be part of the same vPC pair is compatible. For more information, see Virtual Port Channels in Cisco ACI, on page 17.

Figure 14: Supported FEX vPC Topologies



Supported FEX vPC port channel topologies include the following:

- Both VTEP and non-VTEP hypervisors behind a FEX.
- Virtual switches (such as AVS or VDS) connected to two FEXs that are connected to the ACI fabric (vPCs directly connected on physical FEX ports is not supported - a vPC is supported only on port channels).

Note When using GARP as the protocol for notification of IP-to-MAC binding changes to different interfaces on the same FEX, you must set the bridge domain mode to **ARP Flooding** and enable **EP Move Detection Mode**: **GARP-based Detection**, on the **L3 Configuration** page of the bridge domain wizard. This workaround is only required with Generation 1 switches. With Generation 2 switches or later, this is not an issue.

Fibre Channel and FCoE

For Fibre Channel and FCoE configuration information, see the *Cisco APIC Layer 2 Networking Configuration Guide*.

Supporting Fibre Channel over Ethernet Traffic on the Cisco ACI Fabric

Cisco Application Centric Infrastructure (ACI) enables you to configure and manage support for Fibre Channel over Ethernet (FCoE) traffic on the Cisco ACI fabric.

FCoE is a protocol that encapsulates Fibre Channel packets within Ethernet packets, thus enabling storage traffic to move seamlessly between a Fibre Channel SAN and an Ethernet network.

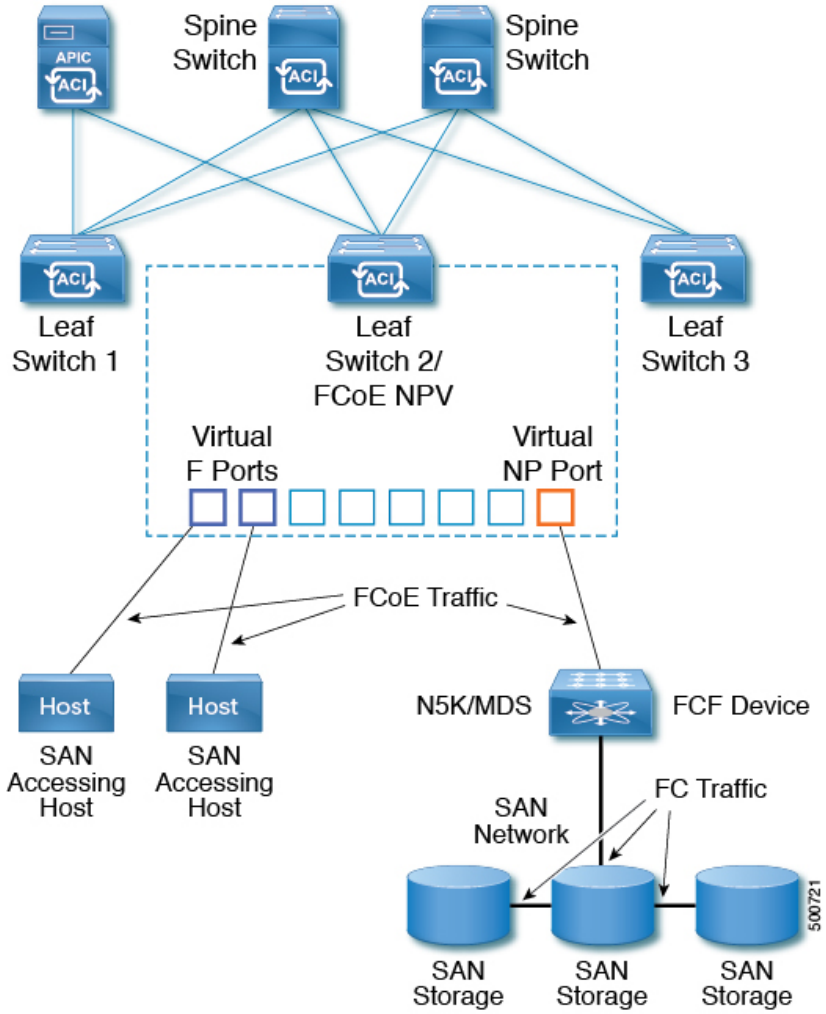
A typical implementation of FCoE protocol support on the Cisco ACI fabric enables hosts located on the Ethernet-based Cisco ACI fabric to communicate with SAN storage devices located on a Fibre Channel network. The hosts are connecting through virtual F ports deployed on an Cisco ACI leaf switch. The SAN storage devices and Fibre Channel network are connected through a Fibre Channel Forwarding (FCF) bridge to the Cisco ACI fabric through a virtual NP port, deployed on the same Cisco ACI leaf switch as is the virtual F port. Virtual NP ports and virtual F ports are also referred to generically as virtual Fibre Channel (vFC) ports.

Note In the FCoE topology, the role of the Cisco ACI leaf switch is to provide a path for FCoE traffic between the locally connected SAN hosts and a locally connected FCF device. The leaf switch does not perform local switching between SAN hosts, and the FCoE traffic is not forwarded to a spine switch.

Topology Supporting FCoE Traffic Through Cisco ACI

The topology of a typical configuration supporting FCoE traffic over the Cisco ACI fabric consists of the following components:

Figure 15: Cisco ACI Topology Supporting FCoE Traffic



- One or more Cisco ACI leaf switches configured through Fibre Channel SAN policies to function as an NPV backbone.
- Selected interfaces on the NPV-configured leaf switches configured to function as virtual F ports, which accommodate FCoE traffic to and from hosts running SAN management or SAN-consuming applications.
- Selected interfaces on the NPV-configured leaf switches configured to function as virtual NP ports, which accommodate FCoE traffic to and from a Fibre Channel Forwarding (FCF) bridge.

The FCF bridge receives Fibre Channel traffic from Fibre Channel links typically connecting SAN storage devices and encapsulates the Fibre Channel packets into FCoE frames for transmission over the Cisco ACI fabric to the SAN management or SAN Data-consuming hosts. It receives FCoE traffic and repackages it back to the Fibre Channel for transmission over the Fibre Channel network.

Note In the above Cisco ACI topology, FCoE traffic support requires direct connections between the hosts and virtual F ports and direct connections between the FCF device and the virtual NP port.

Cisco Application Policy Infrastructure Controller (APIC) servers enable an operator to configure and monitor the FCoE traffic through the Cisco APIC GUI, or NX-OS-style CLI, or through application calls to the REST API.

Topology Supporting FCoE Initialization

In order for FCoE traffic flow to take place as described, you must also set up separate VLAN connectivity over which SAN Hosts broadcast FCoE Initialization protocol (FIP) packets to discover the interfaces enabled as F ports.

vFC Interface Configuration Rules

Whether you set up the vFC network and EPG deployment through the Cisco APIC GUI, NX-OS-style CLI, or the REST API, the following general rules apply across platforms:

- F port mode is the default mode for vFC ports. NP port mode must be specifically configured in the Interface policies.
- The load balancing default mode is for leaf-switch or interface level vFC configuration is src-dst-ox-id.
- One VSAN assignment per bridge domain is supported.
- The allocation mode for VSAN pools and VLAN pools must always be static.
- vFC ports require association with a VSAN domain (also called Fibre Channel domain) that contains VSANs mapped to VLANs.

Fibre Channel Connectivity Overview

Cisco ACI supports Fibre Channel (FC) connectivity on a leaf switch using N-Port Virtualization (NPV) mode. NPV allows the switch to aggregate FC traffic from locally connected host ports (N ports) into a node proxy (NP port) uplink to a core switch.

A switch is in NPV mode after enabling NPV. NPV mode applies to an entire switch. Each end device connected to an NPV mode switch must log in as an N port to use this feature (loop-attached devices are not supported). All links from the edge switches (in NPV mode) to the NPV core switches are established as NP ports (not E ports), which are used for typical inter-switch links.

Note In the FC NPV application, the role of the ACI leaf switch is to provide a path for FC traffic between the locally connected SAN hosts and a locally connected core switch. The leaf switch does not perform local switching between SAN hosts, and the FC traffic is not forwarded to a spine switch.

FC NPV Benefits

FC NPV provides the following:

- Increases the number of hosts that connect to the fabric without adding domain IDs in the fabric. The domain ID of the NPV core switch is shared among multiple NPV switches.
- FC and FCoE hosts connect to SAN fabrics using native FC interfaces.

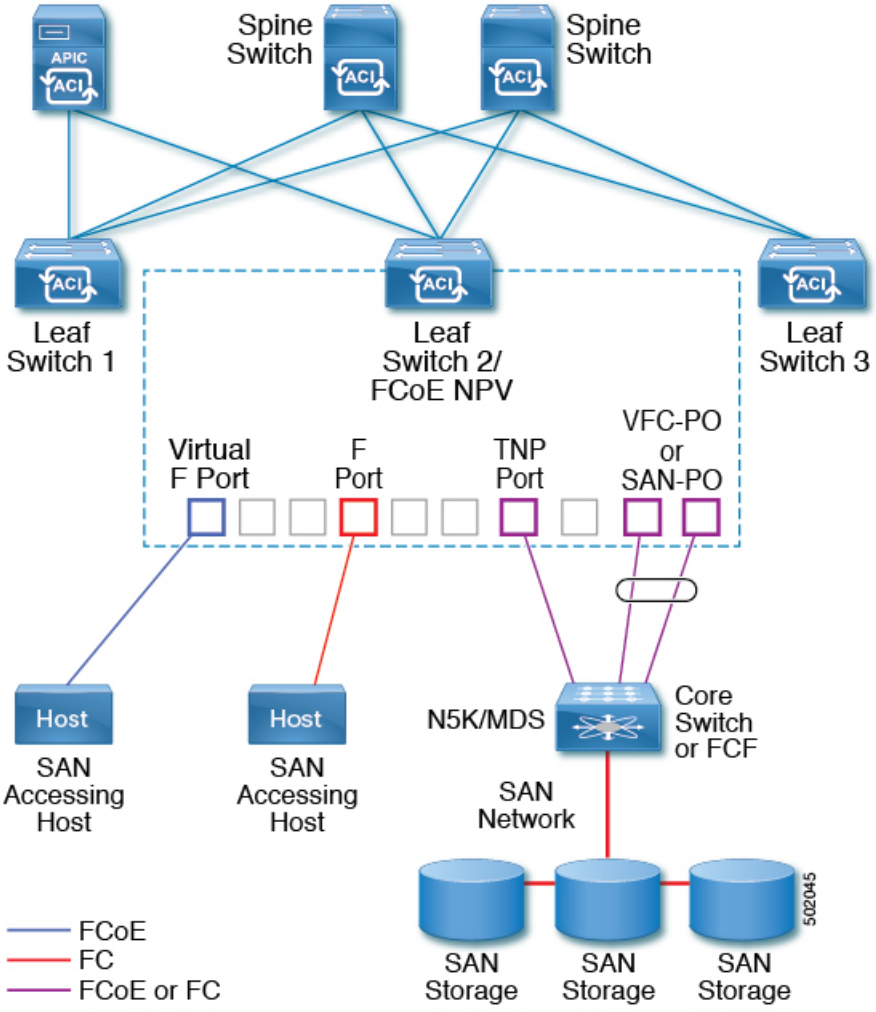
- Automatic traffic mapping for load balancing. For newly added servers connected to NPV, traffic is automatically distributed among the external uplinks based on current traffic loads.
- Static traffic mapping. A server connected to NPV can be statically mapped to an external uplink.

FC NPV Mode

Feature-set fcoe-npv in ACI will be enabled automatically by default when the first FCoE/FC configuration is pushed.

FC Topology

The topology of various configurations supporting FC traffic over the ACI fabric is shown in the following figure:



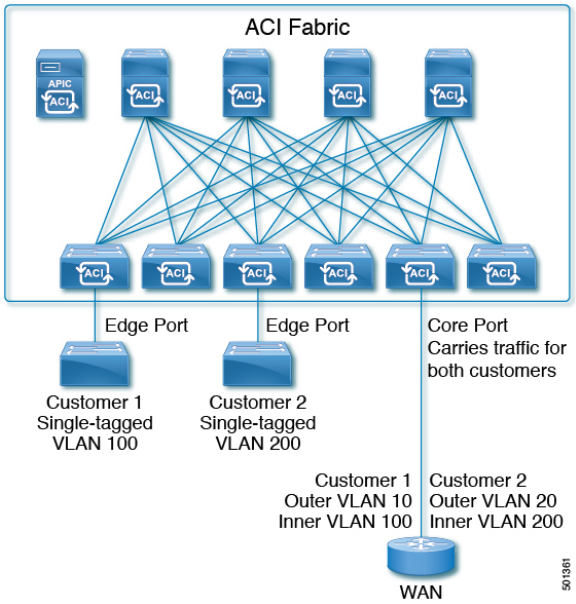
- Server/storage host interfaces on the ACI leaf switch can be configured to function as either native FC ports or as virtual FC (FCoE) ports.
- An uplink interface to a FC core switch can be configured as any of the following port types:

  - native FC NP port
  - SAN-PO NP port
- An uplink interface to a FCF switch can be configured as any of the following port types:
  - virtual (vFC) NP port
  - vFC-PO NP port
- N-Port ID Virtualization (NPIV) is supported and enabled by default, allowing an N port to be assigned multiple N port IDs or Fibre Channel IDs (FCID) over a single link.
- Trunking can be enabled on an NP port to the core switch. Trunking allows a port to support more than one VSAN. When trunk mode is enabled on an NP port, it is referred to as a TNP port.
- Multiple FC NP ports can be combined as a SAN port channel (SAN-PO) to the core switch. Trunking is supported on a SAN port channel.
- FC F ports support 4/16/32 Gbps and auto speed configuration, but 8Gbps is not supported for host interfaces. The default speed is "auto."
- FC NP ports support 4/8/16/32 Gbps and auto speed configuration. The default speed is "auto."
- Multiple FDISC followed by Flogi (nested NPIV) is supported with FC/FCoE host and FC/FCoE NP links.
- An FCoE host behind a FEX is supported over an FCoE NP/uplink.
- Starting in the APIC 4.1(1) release, an FCoE host behind a FEX is supported over the Fibre Channel NP/uplink.
- All FCoE hosts behind one FEX can either be load balanced across multiple vFC and vFC-PO uplinks, or through a single Fibre Channel/SAN port channel uplink.
- SAN boot is supported on a FEX through an FCoE NP/uplink.
- Starting in the APIC 4.1(1) release, SAN boot is also supported over a FC/SAN-PO uplink.
- SAN boot is supported over vPC for FCoE hosts that are connected through FEX.

802.1Q Tunnels

About ACI 802.1Q Tunnels

Figure 16: ACI 802.1Q Tunnels



You can configure 802.1Q tunnels on edge (tunnel) ports to enable point-to-multi-point tunneling of Ethernet frames in the fabric, with Quality of Service (QoS) priority settings. A Dot1q tunnel transports untagged, 802.1Q tagged, and 802.1ad double-tagged frames as-is across the fabric. Each tunnel carries the traffic from a single customer and is associated with a single bridge domain. Cisco Application Centric Infrastructure (ACI) front panel ports can be part of a Dot1q tunnel. Layer 2 switching is done based on the destination MAC (DMAC) and regular MAC learning is done in the tunnel. Edge port Dot1q tunnels are supported on Cisco Nexus 9000 series switches with "EX" or later suffixes in the switch model name.

You can configure multiple 802.1Q tunnels on the same core port to carry double-tagged traffic from multiple customers, each distinguished with an access encapsulation configured for each 802.1Q tunnel. You can also disable MAC address learning on 802.1Q tunnels. Both edge ports and core ports can belong to an 802.1Q tunnel with access encapsulation and disabled MAC address learning. Both edge ports and core ports in Dot1q tunnel are supported on Cisco Nexus 9000 series switches with "FX" or later suffixes in the switch model name.

IGMP and MLD packets can be forwarded through 802.1Q tunnels.

Terms used in this document may be different in the **Cisco Nexus 9000 Series** documents.

Table 1: 802.1Q Tunnel Terminology

ACI Documents	Cisco Nexus 9000 Series Documents
Edge Port	Tunnel Port

ACI Documents	Cisco Nexus 9000 Series Documents
Core Port	Trunk Port

The following guidelines and restrictions apply:

- Layer 2 tunneling of VTP, CDP, LACP, LLDP, and STP protocols is supported with the following restrictions:
  - Link Aggregation Control Protocol (LACP) tunneling functions as expected only with point-to-point tunnels using individual leaf interfaces. It is not supported on port channels (PCs) or virtual port channels (vPCs).
  - CDP and LLDP tunneling with PCs or vPCs is not deterministic; it depends on the link it chooses as the traffic destination.
  - To use VTP for Layer 2 protocol tunneling, CDP must be enabled on the tunnel.
  - STP is not supported in an 802.1Q tunnel bridge domain when Layer 2 protocol tunneling is enabled and the bridge domain is deployed on Dot1q tunnel core ports.
  - Cisco ACI leaf switches react to STP TCN packets by flushing the end points in the tunnel bridge domain and flooding them in the bridge domain.
  - CDP and LLDP tunneling with more than two interfaces flood packets on all interfaces.
  - The destination MAC address of Layer 2 protocol packets tunneled from edge to core ports is rewritten as 01-00-0c-cd-cd-d0 and the destination MAC address of Layer 2 protocol packets tunneled from core to edge ports is rewritten with the standard default MAC address for the protocol.
- If a PC or vPC is the only interface in a Dot1q tunnel and it is deleted and reconfigured, remove the association of the PC/VPC to the Dot1q tunnel and reconfigure it.
- For 802.1Q tunnels deployed on switches that have EX in the product ID, Ethertype combinations of 0x8100+0x8100, 0x8100+0x88a8, 0x88a8+0x8100, and 0x88a8+0x88a8 for the first two VLAN tags are not supported.

  If the tunnels are deployed on a combination of EX and FX or later switches, then this restriction still applies.

  If the tunnels are deployed only on switches that have FX or later in the product ID, then this restriction does not apply.
- For core ports, the Ethertypes for double-tagged frames must be 0x8100 followed by 0x8100.
- You can include multiple edge ports and core ports (even across leaf switches) in a Dot1q tunnel.
- An edge port may only be part of one tunnel, but a core port can belong to multiple Dot1q tunnels.
- Regular EPGs can be deployed on core ports that are used in 802.1Q tunnels.
- L3Outs are not supported on interfaces enabled for Dot1q tunnel.
- FEX interfaces are not supported as members of a Dot1q tunnel.
- Interfaces configured as breakout ports do not support 802.1Q tunnels.
- Interface-level statistics are supported for interfaces in Dot1q tunnel, but statistics at the tunnel level are not supported.

- 802.1Q tunnels are supported across multi-pod fabrics, but not supported across multi-site.

Dynamic Breakout Ports

Configuration of Breakout Ports

Breakout cables are suitable for very short links and offer a cost effective way to connect within racks and across adjacent racks. Breakout enables a 40 Gigabit (Gb) port to be split into four independent and logical 10Gb ports, a 100Gb port to be split into four independent and logical 25Gb ports, or a 400Gb port to be split into four independent and logical 100Gb ports.

Beginning with the 6.1(5)release, a 400G native port can be split into two independent and logical 100Gb ports for NRZ (No Return to Zero) by inserting appropriate 2x100G optics.

You configure breakout on the down links (also known as the access-facing ports or downlink ports) and fabric links of the switches. Fabric links form the connections between the leaf switches and spine switches, or between the tier 1 leaf switches and tier 2 leaf switches for a multi-tier topology.

You can configure breakout ports in the following ways:

- You can use port profiles and selectors. With this method, you configure a breakout leaf port with an leaf interface profile, associate the profile with a switch, and configure the sub-ports.
- Beginning with the Cisco Application Policy Infrastructure Controller (APIC) 6.0(1) release, you can use the **Fabric** > **Access Policies** > **Interface Configuration** workflow.
- You can use the **Fabric** > **Inventory** > ***pod*** > ***leaf_name*** workflow. Beginning with the Cisco APIC 6.0(1) release, the inventory view configuration also uses the interface configuration.

Configuring Port Profiles

Uplink and downlink conversion is supported on Cisco Nexus 9000 series switches with names that end in EX or FX, and later (for example, N9K-C9348GC-FXP or N9K-C93240YC-FX2). A FEX connected to converted downlinks is also supported.

For information about the supported supported Cisco switches, see Port Profile Configuration Summary, on page 33.

When an uplink port is converted to a downlink port, it acquires the same capabilities as any other downlink port.

Restrictions

- Fast Link Failover policies and port profiles are not supported on the same port. If port profile is enabled, Fast Link Failover cannot be enabled or vice versa.
- The last 2 uplink ports of supported leaf switches cannot be converted to downlink ports (they are reserved for uplink connections).
- Dynamic breakouts (both 100Gb and 40Gb) are supported on profiled QSFP ports on the N9K-C93180YC-FX switch. Breakout and port profile are supported together for conversion of uplink

to downlink on ports 49-52. Breakout (both **10g-4x** and **25g-4x** options) is supported on downlink profiled ports.

- The N9K-C9348GC-FXP does not support FEX.
- Breakout is supported only on downlink ports, and not on fabric ports that are connected to other switches.
- A Cisco ACI leaf switch cannot have more than 56 fabric links.
- Reloading a switch after changing a switch's port profile configuration interrupts traffic through the data plane.
- If a port profile is configured on any LEM type and you want to replace that LEM, the replacement LEM type must match the LEM type you removed.

Guidelines

In converting uplinks to downlinks and downlinks to uplinks, consider these guidelines.

Subject	Guideline
Port profile guidelines for N9K-X9400-8D	These guidelines apply for this LEM: • This LEM has 8 ports with a default of 4 downlinks and 4 uplinks. • Port profile conversion is supported on the first 6 ports. • This LEM does not have a port group dependency.
Port profile guidelines for N9K-X9400-16W	These guidelines apply for this LEM: • This LEM has 16 ports, with a default of 12 downlinks and 4 uplinks. • Port profile conversion is supported on the first 6 ports. • This LEM has a port group dependency of 2 ports for port profile conversion. Which means that ports 1-2, 3-4, and 5-6 have a port group dependency. For example, if port 2 is to be converted as an uplink, port 1 should also be converted to an uplink.
Port profile guidelines for N9K-X9400-22L	The 6.1(2) release adds support for this LEM. These guidelines apply: • This LEM has 22 ports with a default of 14 downlinks and 8 uplinks. • Port profile conversion is supported on the first 18 ports. • This LEM has a port group dependency of 4 ports for port profile conversion, except port 9 and 10. This means that ports 1-4, 5-8, 11-14, and 15-18 are part of a port group. For example, if port 2 is to be converted to an uplink, ports 1-4 all need to be converted. • Port 9 and 10 are a port group of 2. If port 10 is to be converted as an uplink, port 9 must become an uplink, also.

Subject	Guideline
LEM Mismatch	On the N9K-C9400-SUP-A, if you have an N9K-X9400-22L LEM with its port profile configured on ports 7-18 and you replace it with another LEM type, like the N9K-X9400-8D or N9K-X9400-16W, the 8D or 16W module will show module status: **LEM type mismatch**. The failure occurs because the port profile conversions do not match. For the 8D and 16W modules, the port profile conversion is ports 1-6 only. To recover the 8D or 16W LEMs from mismatch, remove the port-profile configurations present on all the ports on the previous LEM (i.e., the N9K-X9400-22L LEM). There should not be any port profile configuration. After this process, do a clean reload to recover the LEMs.
Decommissioning nodes with port profiles	If a decommissioned node has the Port Profile feature deployed on it, the port conversions are not removed even after decommissioning the node. It is necessary to manually delete the configurations after decommission, for the ports to return to the default state. To do this, log onto the switch, run the `setup-clean-config.sh` script, and wait for it to run. Then, enter the `reload` command. Optionally, you can specify `-k` with the `setup-clean-config.sh` script to allow the port-profile setting to persist across the reload, making an additional reboot unnecessary. Beginning with 6.0(5), the port-profile setting persists across the reload when running the `setup-clean-config.sh` script with no option, `-k` or `--keep-port-profile`. To manually delete the configuration, run the `setup-clean-config.sh` script with `-d` or `--delete-profiles`.
Maximum uplink port limit	When the maximum uplink port limit is reached and ports 25 and 27 are converted from uplink to downlink and back to uplink on Cisco 93180LC-EX switches: On Cisco N9K-93180LC-EX switches, ports 25 and 27 are the original uplink ports. Using the port profile, if you convert port 25 and 27 to downlink ports, ports 29, 30, 31, and 32 are still available as four original uplink ports. Because of the threshold on the number of ports (which is maximum of 12 ports) that can be converted, you can convert 8 more downlink ports to uplink ports. For example, ports 1, 3, 5, 7, 9, 13, 15, 17 are converted to uplink ports and ports 29, 30, 31 and 32 are the 4 original uplink ports (the maximum uplink port limit on Cisco 93180LC-EX switches). When the switch is in this state and if the port profile configuration is deleted on ports 25 and 27, ports 25 and 27 are converted back to uplink ports, but there are already 12 uplink ports on the switch (as mentioned earlier). To accommodate ports 25 and 27 as uplink ports, 2 random ports from the port range 1, 3, 5, 7, 9, 13, 15, 17 are denied the uplink conversion and this situation cannot be controlled by the user. Therefore, it is mandatory to clear all the faults before reloading the leaf node to avoid any unexpected behavior regarding the port type. It should be noted that if a node is reloaded without clearing the port profile faults, especially when there is a fault related to limit-exceed, the port might not be in an expected operational state.

Breakout Limitations

Switch	Releases	Limitations
N9K-C93180LC-EX	Cisco APIC 3.1(1) and later	• 40Gb and 100Gb dynamic breakouts are supported on ports 1 through 24 on odd numbered ports. • When the top ports (odd ports) are broken out, then the bottom ports (even ports) are error disabled. • Port profiles and breakouts are not supported on the same port. However, you can apply a port profile to convert a fabric port to a downlink, and then apply a breakout configuration.
N9K-C9336C-FX2-E	Cisco APIC 5.2(4) and later	• 40Gb and 100Gb dynamic breakouts are supported on ports 1 through 34. • A port profile cannot be applied to a port with breakout enabled. However, you can apply a port profile to convert a fabric port to a downlink, and then apply a breakout configuration. • All 34 ports can be configured as breakout ports. • If you want to apply a breakout configuration on 34 ports, you must configure a port profile on the ports to have 34 downlink ports, then you must reboot the leaf switch. • If you apply a breakout configuration to a leaf switch for multiple ports at the same time, it can take up to 10 minutes for the hardware of 34 ports to be programmed. The ports remain down until the programming completes. The delay can occur for a new configuration, after a clean reboot, or during switch discovery.

Switch	Releases	Limitations
N9K-C9336C-FX2	Cisco APIC 4.2(4) and later	• 40Gb and 100Gb dynamic breakouts are supported on ports 1 through 34. • A port profile cannot be applied to a port with breakout enabled. However, you can apply a port profile to convert a fabric port to a downlink, and then apply a breakout configuration. • All 34 ports can be configured as breakout ports. • If you want to apply a breakout configuration on 34 ports, you must configure a port profile on the ports to have 34 downlink ports, then you must reboot the leaf switch. • If you apply a breakout configuration to a leaf switch for multiple ports at the same time, it can take up to 10 minutes for the hardware of 34 ports to be programmed. The ports remain down until the programming completes. The delay can occur for a new configuration, after a clean reboot, or during switch discovery.
N9K-C9336C-FX2	Cisco APIC 3.2(1) up through, but not including, 4.2(4)	• 40Gb and 100Gb dynamic breakouts are supported on ports 1 through 30. • Port profiles and breakouts are not supported on the same port. However, you can apply a port profile to convert a fabric port to a downlink, and then apply a breakout configuration. • A maximum of 20 ports can be configured as breakout ports.

Switch	Releases	Limitations
N9K-C93180YC-FX	Cisco APIC 3.2(1) and later	• 40Gb and 100Gb dynamic breakouts are supported on ports 49 though 52, when they are on profiled QSFP ports. To use them for dynamic breakout, perform these steps: • Convert ports 49-52 to front panel ports (downlinks). • Perform a port-profile reload, using one of these methods: • In the Cisco APIC GUI, navigate to **Fabric** > **Inventory** > **Pod** > **Leaf**, right-click **Chassis** and choose **Reload**. • In the iBash CLI, enter the **reload** command. • Apply breakouts on the profiled ports 49-52. • Ports 53 and 54 do not support either port profiles or breakouts.
N9K-C93240YC-FX2	Cisco APIC 4.0(1) and later	Breakout is not supported on converted downlinks.

Port Profile Configuration Summary

The following table summarizes supported uplinks and downlinks for the switches that support port profile conversions from uplink to downlink and downlink to uplink.

Switch Model	Default Links	Max Uplinks (Fabric Ports)	Max Downlinks (Server Ports)	Release Supported
N9K-C9348GC-FXP[1] N9K-C9348GC-FX3	48 x 100M/1G BASE-T downlinks 4 x 10/25 Gbps SFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	48 x 100M/1G BASE-T downlinks 4 x 10/25 Gbps SFP28 uplinks 2 x 40/100 Gbps QSFP28 uplinks	Same as default port configuration	3.1(1) 6.0(5)

Switch Model	Default Links	Max Uplinks (Fabric Ports)	Max Downlinks (Server Ports)	Release Supported
N9K-C93180LC-EX	24 x 40 Gbps QSFP28 downlinks (ports 1-24) 2 x 40/100 Gbps QSFP28 uplinks (ports 25, 27) 4 x 40/100 Gbps QSFP28 uplinks (ports 29-32) Or 12 x 100 Gbps QSFP28 downlinks (odd number ports from 1-24) 2 x 40/100 Gbps QSFP28 uplinks (ports 25, 27) 4 x 40/100 Gbps QSFP28 uplinks (ports 29-32)	18 x 40 Gbps QSFP28 downlinks (from 1-24) 6 x 40 Gbps QSFP28 uplinks(from 1-24) 2 x 40/100 Gbps QSFP28 uplinks(25, 27) 4 x 40/100 Gbps QSFP28 uplinks(29-32) Or 6 x 100 Gbps QSFP28 downlinks(odd number from 1-24) 6 x 100 Gbps QSFP28 uplinks(odd number from 1-24) 2 x 40/100 Gbps QSFP28 uplinks(25, 27) 4 x 40/100 Gbps QSFP28 uplinks(29-32)	24 x 40 Gbps QSFP28 downlinks(1-24) 2 x 40/100 Gbps QSFP28 downlinks(25, 27) 4 x 40/100 Gbps QSFP28 uplinks(29-32) Or 12 x 100 Gbps QSFP28 downlinks(odd number from 1-24) 2 x 40/100 Gbps QSFP28 downlinks (25, 27) 4 x 40/100 Gbps QSFP28 uplinks(29-32)	3.1(1)
N9K-C93180YC-EX N9K-C93180YC-FX	48 x 10/25 Gbps fiber downlinks 6 x 40/100 Gbps QSFP28 uplinks	Same as default port configuration	48 x 10/25 Gbps fiber downlinks 4 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	3.1(1)
		48 x 10/25 Gbps fiber uplinks 6 x 40/100 Gbps QSFP28 uplinks		4.0(1)
N9K-C93180YC-FX3	48 x 10/25 Gbps fiber downlinks 6 x 40/100 Gbps QSFP28 uplinks	18 x 10/25 Gbps fiber downlinks 30 x 10/25 Gbps fiber uplinks 6 x 40/100 Gbps QSFP28 uplinks	48 x 10/25 Gbps fiber downlinks 4 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	5.1(3)
		48 x 10/25 Gbps fiber uplinks 6 x 40/100 Gbps QSFP28 uplinks		6.1(5)

Switch Model	Default Links	Max Uplinks (Fabric Ports)	Max Downlinks (Server Ports)	Release Supported
N9K-C93108TC-EX[2] N9K-C93108TC-FX[2] N9K-C93108TC-FX3	48 x 10GBASE-T downlinks 6 x 40/100 Gbps QSFP28 uplinks	Same as default port configuration	48 x 10/25 Gbps fiber downlinks 4 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	3.1(1)
				4.0(1)
				5.1(3)
N9K-C9336C-FX2	30 x 40/100 Gbps QSFP28 downlinks 6 x 40/100 Gbps QSFP28 uplinks	18 x 40/100 Gbps QSFP28 downlinks 18 x 40/100 Gbps QSFP28 uplinks	Same as default port configuration	3.2(1)
		18 x 40/100 Gbps QSFP28 downlinks 18 x 40/100 Gbps QSFP28 uplinks	34 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	3.2(3)
		36 x 40/100 Gbps QSFP28 uplinks	34 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	4.1(1)
N9K-C9336C-FX2-E	30 x 40/100 Gbps QSFP28 downlinks 6 x 40/100 Gbps QSFP28 uplinks	36 x 40/100 Gbps QSFP28 uplinks	34 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	5.2(4)
N9K-93240YC-FX2	48 x 10/25 Gbps fiber downlinks 12 x 40/100 Gbps QSFP28 uplinks	Same as default port configuration	48 x 10/25 Gbps fiber downlinks 10 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	4.0(1)
		48 x 10/25 Gbps fiber uplinks 12 x 40/100 Gbps QSFP28 uplinks		4.1(1)
N9K-C93216TC-FX2	96 x 10G BASE-T downlinks 12 x 40/100 Gbps QSFP28 uplinks	Same as default port configuration	96 x 10G BASE-T downlinks 10 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	4.1(2)

Switch Model	Default Links	Max Uplinks (Fabric Ports)	Max Downlinks (Server Ports)	Release Supported
N9K-C93360YC-FX2	96 x 10/25 Gbps SFP28 downlinks 12 x 40/100 Gbps QSFP28 uplinks	44 x 10/25Gbps SFP28 downlinks 52 x 10/25Gbps SFP28 uplinks 12 x 40/100Gbps QSFP28 uplinks	96 x 10/25 Gbps SFP28 downlinks 10 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	4.1(2)
N9K-C93600CD-GX	28 x 40/100 Gbps QSFP28 downlinks (ports 1-28) 8 x 40/100/400 Gbps QSFP-DD uplinks (ports 29-36)	28 x 40/100 Gbps QSFP28 uplinks 8 x 40/100/400 Gbps QSFP-DD uplinks	28 x 40/100 Gbps QSFP28 downlinks 6 x 40/100/400 Gbps QSFP-DD downlinks 2 x 40/100/400 Gbps QSFP-DD uplinks	4.2(2)
N9K-C9364C-GX	48 x 40/100 Gbps QSFP28 downlinks (ports 1-48) 16 x 40/100 Gbps QSFP28 uplinks (ports 49-64)	56 x 40/100 Gbps QSFP28 uplinks	62 x 40/100 Gbps QSFP28 downlinks 2 x 40/100 Gbps QSFP28 uplinks	4.2(3)
N9K-C9316D-GX	12 x 40/100/400 Gbps QSFP-DD downlinks (ports 1-12) 4 x 40/100/400 Gbps QSFP-DD uplinks (ports 13-16)	16 x 40/100/400 Gbps QSFP-DD uplinks	14 x 40/100/400 Gbps QSFP-DD downlinks	5.1(4)
N9K-C9332D-GX2B	2 x 1/10 Gbps SFP+ downlinks (ports 33-34) 24 x 40/100/400 Gbps QSFP-DD downlinks (ports 1-24) 8 x 40/100/400 Gbps QSFP-DD uplinks (ports 25-32)	2 x 1/10 Gbps SFP+ downlinks 32 x 40/100/400 Gbps QSFP-DD uplinks	2 x 1/10 Gbps SFP+ downlinks 30 x 40/100/400 Gbps QSFP-DD downlinks 2 x 40/100/400 Gbps QSFP-DD uplinks	5.2(3)

Switch Model	Default Links	Max Uplinks (Fabric Ports)	Max Downlinks (Server Ports)	Release Supported
N9K-C9348D-GX2A	2 x 1/10 Gbps SFP+ downlinks (ports 49-50) 36 x 40/100/400 Gbps QSFP-DD downlinks (ports 1-36) 12 x 40/100/400 Gbps QSFP-DD uplinks (ports 37-48)	2 x 1/10 Gbps SFP+ downlinks 48 x 40/100/400 Gbps QSFP-DD uplinks	2 x 1/10 Gbps SFP+ downlinks 46 x 40/100/400 Gbps QSFP-DD downlinks 2 x 40/100/400 Gbps QSFP-DD uplinks	5.2(5)
N9K-C9364D-GX2A	2 x 1/10 Gbps SFP+ downlinks (ports 65-66) 48 x 40/100/400 Gbps QSFP-DD downlinks (ports 1-48) 16 x 40/100/400 Gbps QSFP-DD uplinks (ports 49-64)	2 x 1/10 Gbps SFP+ downlinks 56 x 40/100/400 Gbps QSFP-DD uplinks	2 x 1/10 Gbps SFP+ downlinks 62 x 40/100/400 Gbps QSFP-DD downlinks 2 x 40/100/400 Gbps QSFP-DD uplinks	5.2(5)
N9K-C9408 with N9K-X9400-8D[3]	6 x 40/100/400 Gbps QSFP-DD downlinks 2 x 40/100/400 Gbps QSFP-DD uplinks	8 x 40/100/400 Gbps QSFP-DD uplinks	Same as default port configuration	6.0(2)
N9K-C9408 with N9K-X9400-16W[3]	12 x 100/200 Gbps QSFP56 downlinks 4 x 100/200 Gbps QSFP56 uplinks	6 x 100/200 Gbps QSFP56 uplinks (ports 1-6) 6 x 100/200 Gbps QSFP56 downlinks (ports 7-12) 4 x 100/200 Gbps QSFP56 uplinks (ports 13-16)	Same as default port configuration	6.0(2)[4]

1 Does not support FEX.

2 Only uplink to downlink conversion is supported.

3 Only ports 1 through 6 support port profile conversion.

4 The 6.0(2) release does not support 200 Gbps.

Port Tracking Policy for Fabric Port Failure Detection

Fabric port failure detection can be enabled in the port tracking system settings. The port tracking policy monitors the status of fabric ports between leaf switches and spine switches, and ports between tier-1 leaf switches and tier-2 leaf switches. When an enabled port tracking policy is triggered, the leaf switches take down all access interfaces on the switch that have EPGs deployed on them.

If you enabled the **Include APIC ports when port tracking is triggered** option, port tracking disables Cisco Application Policy Infrastructure Controller (APIC) ports when the leaf switch loses connectivity to all fabric ports (that is, there are 0 fabric ports). Enable this feature only if the Cisco APICs are dual- or multihomed to the fabric. Bringing down the Cisco APIC ports helps in switching over to the secondary port in the case of a dual-homed Cisco APIC.

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

The port tracking policy specifies the number of fabric port connections that trigger the policy, and a delay timer for bringing the leaf switch access ports back up after the number of specified fabric ports is exceeded.

The following example illustrates how a port tracking policy behaves:

- The port tracking policy specifies that the threshold of active fabric port connections each leaf switch that triggers the policy is 2.
- The port tracking policy triggers when the number of active fabric port connections from the leaf switch to the spine switches drops to 2.
- Each leaf switch monitors its fabric port connections and triggers the port tracking policy according to the threshold specified in the policy.
- When the fabric port connections come back up, the leaf switch waits for the delay timer to expire before bringing its access ports back up. This gives the fabric time to reconverge before allowing traffic to resume on leaf switch access ports. Large fabrics may need the delay timer to be set for a longer time.

Note Use caution when configuring this policy. If the port tracking setting for the number of active spine ports that triggers port tracking is too high, all leaf switch access ports will be brought down.

Q-in-Q Encapsulation Mapping for EPGs

Using Cisco Application Policy Infrastructure Controller (APIC), you can map double-tagged VLAN traffic ingressing on a regular interface, PC, or vPC to an EPG. When this feature is enabled, when double-tagged traffic enters the network for an EPG, both tags are processed individually in the fabric and restored to double-tags when egressing the Cisco Application Centric Infrastructure (ACI) switch. Ingressing single-tagged and untagged traffic is dropped.

The following guidelines and limitations apply:

- This feature is supported on Cisco Nexus 9300-FX, 9300-FX2, and 9300-FX3 platform switches.

- Both the outer and inner tag must be of EtherType 0x8100.
- MAC learning and routing are based on the EPG port, sclass, and VRF instance, not on the access encapsulations.
- QoS priority settings are supported, derived from the outer tag on ingress, and rewritten to both tags on egress.
- EPGs can simultaneously be associated with other interfaces on a leaf switch, that are configured for single-tagged VLANs.
- Service graphs are supported for provider and consumer EPGs that are mapped to Q-in-Q encapsulated interfaces. You can insert service graphs, as long as the ingress and egress traffic on the service nodes is in single-tagged encapsulated frames.
- When vPC ports are enabled for Q-in-Q encapsulation mode, VLAN consistency checks are not performed.

The following features and options are not supported with this feature:

- Per-port VLAN feature
- FEX connections
- Mixed mode

  For example, an interface in Q-in-Q encapsulation mode can have a static path binding to an EPG with double-tagged encapsulation only, not with regular VLAN encapsulation.
- STP and the "Flood in Encapsulation" option
- Untagged and 802.1p mode
- Multi-pod and Multi-Site
- Legacy bridge domain
- L2Out and L3Out connections
- VMM integration
- Changing a port mode from routed to Q-in-Q encapsulation mode
- Per-VLAN mis-cabling protocol on ports in Q-in-Q encapsulation mode

Layer 2 Multicast

About Cisco APIC and IGMP Snooping

IGMP snooping is the process of listening to Internet Group Management Protocol (IGMP) network traffic. The feature allows a network switch to listen in on the IGMP conversation between hosts and routers and filter multicasts links that do not need them, thus controlling which ports receive specific multicast traffic.

Cisco APIC provides support for the full IGMP snooping feature included on a traditional switch such as the N9000 standalone.

- Policy-based IGMP snooping configuration per bridge domain

APIC enables you to configure a policy in which you enable, disable, or customize the properties of IGMP Snooping on a per bridge-domain basis. You can then apply that policy to one or multiple bridge domains.

- Static port group implementation

  IGMP static port grouping enables you to pre-provision ports, already statically-assigned to an application EPG, as the switch ports to receive and process IGMP multicast traffic. This pre-provisioning prevents the join latency which normally occurs when the IGMP snooping stack learns ports dynamically.

  Static group membership can be pre-provisioned only on static ports (also called, *static-binding ports*) assigned to an application EPG.

- Access group configuration for application EPGs

  An “access-group” is used to control what streams can be joined behind a given port.

  An access-group configuration can be applied on interfaces that are statically assigned to an application EPG in order to ensure that the configuration can be applied on ports that will actually belong to the that EPG.

  Only Route-map-based access groups are allowed.

Note You can use **vzAny** to enable protocols such as IGMP Snooping for all the EPGs in a VRF. For more information about **vzAny**, see Use vzAny to Automatically Apply Communication Rules to all EPGs in a VRF.

To use **vzAny**, navigate to **Tenants** > ***tenant-name*** > **Networking** > **VRFs** > ***vrf-name*** > **EPG Collection for VRF**.

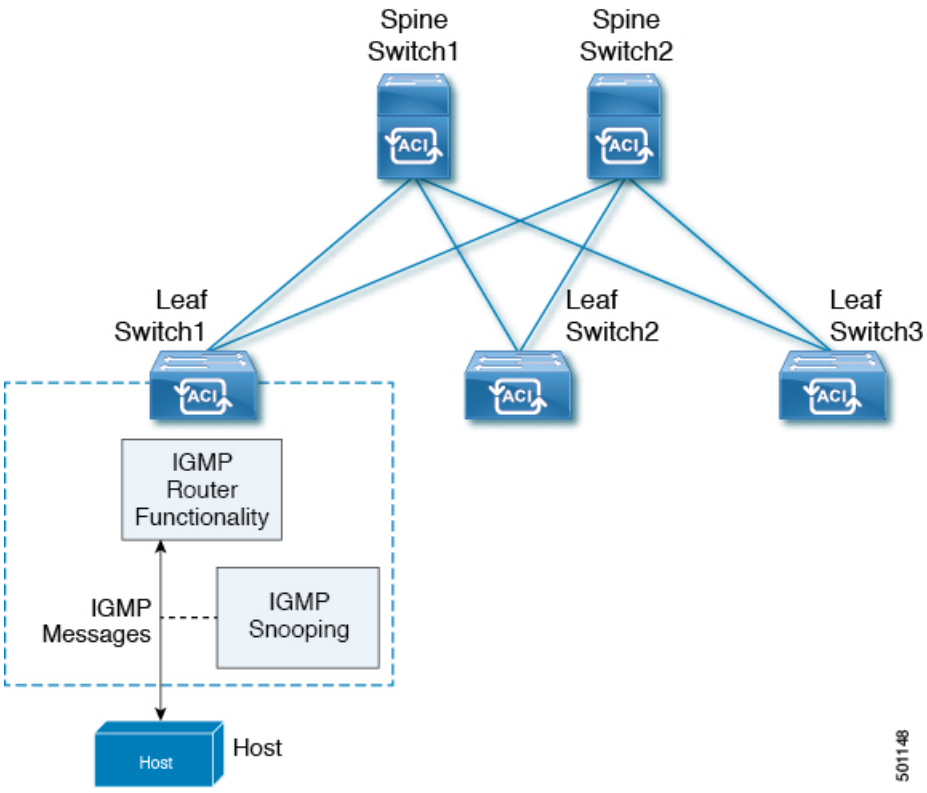
How IGMP Snooping is Implemented in the ACI Fabric

Note We recommend that you do not disable IGMP snooping on bridge domains. If you disable IGMP snooping, you may see reduced multicast performance because of excessive false flooding within the bridge domain.

IGMP snooping software examines IP multicast traffic within a bridge domain to discover the ports where interested receivers reside. Using the port information, IGMP snooping can reduce bandwidth consumption in a multi-access bridge domain environment to avoid flooding the entire bridge domain. By default, IGMP snooping is enabled on the bridge domain.

This figure shows the IGMP routing functions and IGMP snooping functions both contained on an ACI leaf switch with connectivity to a host. The IGMP snooping feature snoops the IGMP membership reports, and leaves messages and forwards them only when necessary to the IGMP router function.

Figure 17: IGMP Snooping function



IGMP snooping operates upon IGMPv1, IGMPv2, and IGMPv3 control plane packets where Layer 3 control plane packets are intercepted and influence the Layer 2 forwarding behavior.

IGMP snooping has the following proprietary features:

- Source filtering that allows forwarding of multicast packets based on destination and source IP addresses
- Multicast forwarding based on IP addresses rather than the MAC address
- Multicast forwarding alternately based on the MAC address

The ACI fabric supports IGMP snooping only in proxy-reporting mode, in accordance with the guidelines provided in Section 2.1.1, "IGMP Forwarding Rules," in RFC 4541:

> IGMP networks may also include devices that implement "proxy-reporting", in which reports received from downstream hosts are summarized and used to build internal membership states. Such proxy-reporting devices may use the all-zeros IP Source-Address when forwarding any summarized reports upstream. For this reason, IGMP membership reports received by the snooping switch must not be rejected because the source IP address is set to 0.0.0.0.

As a result, the ACI fabric will send IGMP reports with the source IP address of 0.0.0.0.

Note For more information about IGMP snooping, see RFC 4541.

Virtualization Support

You can define multiple virtual routing and forwarding (VRF) instances for IGMP snooping.

On leaf switches, you can use the **show** commands with a VRF argument to provide a context for the information displayed. The default VRF is used if no VRF argument is supplied.

The APIC IGMP Snooping Function, IGMPv1, IGMPv2, and the Fast Leave Feature

Both IGMPv1 and IGMPv2 support membership report suppression, which means that if two hosts on the same subnet want to receive multicast data for the same group, the host that receives a member report from the other host suppresses sending its report. Membership report suppression occurs for hosts that share a port.

If no more than one host is attached to each switch port, you can configure the fast leave feature in IGMPv2. The fast leave feature does not send last member query messages to hosts. As soon as APIC receives an IGMP leave message, the software stops forwarding multicast data to that port.

IGMPv1 does not provide an explicit IGMP leave message, so the APIC IGMP snooping function must rely on the membership message timeout to indicate that no hosts remain that want to receive multicast data for a particular group.

Note The IGMP snooping function ignores the configuration of the last member query interval when you enable the fast leave feature because it does not check for remaining hosts.

The APIC IGMP Snooping Function and IGMPv3

The IGMPv3 snooping function in APIC supports full IGMPv3 snooping, which provides constrained flooding based on the (S, G) information in the IGMPv3 reports. This source-based filtering enables the device to constrain multicast traffic to a set of ports based on the source that sends traffic to the multicast group.

By default, the IGMP snooping function tracks hosts on each VLAN port in the bridge domain. The explicit tracking feature provides a fast leave mechanism. Because every IGMPv3 host sends membership reports, report suppression limits the amount of traffic that the device sends to other multicast-capable routers. When report suppression is enabled, and no IGMPv1 or IGMPv2 hosts requested the same group, the IGMP snooping function provides proxy reporting. The proxy feature builds the group state from membership reports from the downstream hosts and generates membership reports in response to queries from upstream queriers.

Even though the IGMPv3 membership reports provide a full accounting of group members in a bridge domain, when the last host leaves, the software sends a membership query. You can configure the parameter last member query interval. If no host responds before the timeout, the IGMP snooping function removes the group state.

Cisco APIC and the IGMP Snooping Querier Function

When PIM is not enabled on an interface because the multicast traffic does not need to be routed, you must configure an IGMP snooping querier function to send membership queries. In APIC, within the IGMP Snoop policy, you define the querier in a bridge domain that contains multicast sources and receivers but no other active querier.

Cisco ACI has by default, IGMP snooping enabled. Additionally, if the Bridge Domain subnet control has "querier IP" selected, then the leaf switch behaves as a querier and starts sending query packets. Querier on the ACI leaf switch must be enabled when the segments do not have an explicit multicast router (PIM is not enabled). On the Bridge Domain where the querier is configured, the IP address used must be from the same subnet where the multicast hosts are configured.

Note The IP address for the querier should not be a broadcast IP address, multicast IP address, or 0 (0.0.0.0).

When an IGMP snooping querier is enabled, it sends out periodic IGMP queries that trigger IGMP report messages from hosts that want to receive IP multicast traffic. IGMP snooping listens to these IGMP reports to establish appropriate forwarding.

The IGMP snooping querier performs querier election as described in RFC 2236. Querier election occurs in the following configurations:

- When there are multiple switch queriers configured with the same subnet on the same VLAN on different switches.
- When the configured switch querier is in the same subnet as with other Layer 3 SVI queriers.

Fabric Secure Mode

Fabric secure mode prevents parties with physical access to the fabric equipment from adding a switch or APIC controller to the fabric without manual authorization by an administrator. Starting with release 1.2(1x), the firmware checks that switches and controllers in the fabric have valid serial numbers associated with a valid Cisco digitally signed certificate. This validation is performed upon upgrade to this release or during an initial installation of the fabric. The default setting for this feature is permissive mode; an existing fabric continues to run as it has after an upgrade to release 1.2(1) or later. An administrator with fabric-wide access rights must enable strict mode. The following table summarizes the two modes of operation:

Permissive Mode (default)	Strict Mode
Allows an existing fabric to operate normally even though one or more switches have an invalid certificate.	Only switches with a valid Cisco serial number and SSL certificate are allowed.
Does not enforce serial number based authorization .	Enforces serial number authorization.
Allows auto-discovered controllers and switches to join the fabric without enforcing serial number authorization.	Requires an administrator to manually authorize controllers and switches to join the fabric.

Configuring Fast Link Failover Policy

Fast Link Failover policy is applicable to uplinks on switch models with -EX, -FX, and -FX2 suffixes. It efficiently load balances the traffic based on the uplink MAC status. With this functionality, the switch performs Layer 2 or Layer 3 lookup and it provides an output Layer 2 interface (uplinks) based on the packet hash algorithm by considering the uplink status. This functionality reduces the data traffic convergence to less than 200 milliseconds.

See the following limitations on configuring Fast Link Failover:

- Fast Link Failover and port profiles are not supported on the same interface. If port profile is enabled, Fast Link Failover cannot be enabled or vice versa.
- Configuring remote leaf does not work with Fast Link Failover. In this case, Fast Link Failover policies will not work and no fault will be generated.
- When Fast Link Failover policy is enabled, configuring SPAN on individual uplinks will not work. No fault will be generated while attempting to enable SPAN on individual uplinks but Fast Link Failover policy can be enabled on all uplinks together or it can be enabled on an individual downlink.

Note Fast Link Failover is located under **Fabric** > **Access Policies** > **Policies** > **Switch** > **Fast Link Failover**.

About Port Security and ACI

The port security feature protects the ACI fabric from being flooded with unknown MAC addresses by limiting the number of MAC addresses learned per port. The port security feature support is available for physical ports, port channels, and virtual port channels.

Port Security and Learning Behavior

For non-vPC ports or port channels, whenever a learn event comes for a new endpoint, a verification is made to see if a new learn is allowed. If the corresponding interface has a port security policy not configured or disabled, the endpoint learning behavior is unchanged with what is supported. If the policy is enabled and the limit is reached, the current supported action is as follows:

- Learn the endpoint and install it in the hardware with a drop action.
- Silently discard the learn.

If the limit is not reached, the endpoint is learned and a verification is made to see if the limit is reached because of this new endpoint. If the limit is reached, and the learn disable action is configured, learning will be disabled in the hardware on that interface (on the physical interface or on a port channel or vPC). If the limit is reached and the learn disable action is not configured, the endpoint will be installed in hardware with a drop action. Such endpoints are aged normally like any other endpoints.

When the limit is reached for the first time, the operational state of the port security policy object is updated to reflect it. A static rule is defined to raise a fault so that the user is alerted. A syslog is also raised when the limit is reached.

In case of vPC, when the MAC limit is reached, the peer leaf switch is also notified so learning can be disabled on the peer. As the vPC peer can be rebooted any time or vPC legs can become unoperational or restart, this state will be reconciled with the peer so vPC peers do not go out of sync with this state. If they get out of sync, there can be a situation where learning is enabled on one leg and disabled on the other leg.

By default, once the limit is reached and learning is disabled, it will be automatically re-enabled after the default timeout value of 60 seconds.

Protect Mode

The protect mode prevents further port security violations from occurring. Once the MAC limit exceeds the maximum configured value on a port, all traffic from excess MAC addresses will be dropped and further learning is disabled.

Port Security at Port Level

In the APIC, the user can configure the port security on switch ports. Once the MAC limit has exceeded the maximum configured value on a port, all traffic from the exceeded MAC addresses is forwarded. The following attributes are supported:

- **Port Security Timeout**—The current supported range for the timeout value is from 60 to 3600 seconds.
- **Violation Action**—The violation action is available in protect mode. In the protect mode, MAC learning is disabled and MAC addresses are not added to the CAM table. Mac learning is re-enabled after the configured timeout value.
- **Maximum Endpoints**—The current supported range for the maximum endpoints configured value is from 0 to 12000. If the maximum endpoints value is 0, the port security policy is disabled on that port.

Port Security Guidelines and Restrictions

The guidelines and restrictions are as follows:

- Port security is available per port.
- Port security is supported for physical ports, port channels, and virtual port channels (vPCs).
- Static and dynamic MAC addresses are supported.
- MAC address moves are supported from secured to unsecured ports and from unsecured ports to secured ports.
- The MAC address limit is enforced only on the MAC address and is not enforced on a MAC and IP address.
- Port security is not supported with the Fabric Extender (FEX).

About First Hop Security

First-Hop Security (FHS) features enable a better IPv4 and IPv6 link security and management over the layer 2 links. In a service provider environment, these features closely control address assignment and derived operations, such as Duplicate Address Detection (DAD) and Address Resolution (AR).

The following supported FHS features secure the protocols and help build a secure endpoint database on the fabric leaf switches, that are used to mitigate security threats such as MIM attacks and IP thefts:

- ARP Inspection—allows a network administrator to intercept, log, and discard ARP packets with invalid MAC address to IP address bindings.
- ND Inspection—learns and secures bindings for stateless autoconfiguration addresses in Layer 2 neighbor tables.
- DHCP Inspection—validates DHCP messages received from untrusted sources and filters out invalid messages.
- RA Guard—allows the network administrator to block or reject unwanted or rogue router advertisement (RA) guard messages.
- IPv4 and IPv6 Source Guard—blocks any data traffic from an unknown source.
- Trust Control—a trusted source is a device that is under your administrative control. These devices include the switches, routers, and servers in the Fabric. Any device beyond the firewall or outside the network is an untrusted source. Generally, host ports are treated as untrusted sources.

FHS features provide the following security measures:

- **Role Enforcement**—Prevents untrusted hosts from sending messages that are out the scope of their role.
- **Binding Enforcement**—Prevents address theft.
- **DoS Attack Mitigations**—Prevents malicious end-points to grow the end-point database to the point where the database could stop providing operation services.
- **Proxy Services**—Provides some proxy-services to increase the efficiency of address resolution.

FHS features are enabled on a per tenant bridge domain (BD) basis. As the bridge domain, may be deployed on a single or across multiple leaf switches, the FHS threat control and mitigation mechanisms cater to a single switch and multiple switch scenarios.

Beginning with Cisco APIC release 6.0(2), FHS is supported on the VMware DVS VMM domain. If you need to implement FHS within an EPG, enable intra EPG isolation. If intra EPG isolation is not enabled, then, the endpoints within the same VMware ESX port-group can bypass FHS. If you do not enable intra EPG isolation, FHS features still take effect for endpoints that are in different port-groups, for instance, FHS can prevent a compromised VM from poisoning the ARP table of another VM in a different port-group.

About MACsec

MACsec is an IEEE 802.1AE standards based Layer 2 hop-by-hop encryption that provides data confidentiality and integrity for media access independent protocols.

MACsec, provides MAC-layer encryption over wired networks by using out-of-band methods for encryption keying. The MACsec Key Agreement (MKA) Protocol provides the required session keys and manages the required encryption keys.

The 802.1AE encryption with MKA is supported on all types of links, that is, host facing links (links between network access devices and endpoint devices such as a PC or IP phone), or links connected to other switches or routers.

MACsec encrypts the entire data except for the Source and Destination MAC addresses of an Ethernet packet. The user also has the option to skip encryption up to 50 bytes after the source and destination MAC address.

To provide MACsec services over the WAN or Metro Ethernet, service providers offer Layer 2 transparent services such as E-Line or E-LAN using various transport layer protocols such as Ethernet over Multiprotocol Label Switching (EoMPLS) and L2TPv3.

The packet body in an EAP-over-LAN (EAPOL) Protocol Data Unit (PDU) is referred to as a MACsec Key Agreement PDU (MKPDU). When no MKPDU is received from a participants after 3 hearbeats (each hearbeat is of 2 seconds), peers are deleted from the live peer list. For example, if a client disconnects, the participant on the switch continues to operate MKA until 3 heartbeats have elapsed after the last MKPDU is received from the client.

APIC Fabric MACsec

The APIC will be responsible for the MACsec keychain distribution to all the nodes in a Pod or to particular ports on a node. Below are the supported MACsec keychain and MACsec policy distribution supported by the APIC.

- A single user provided keychain and policy per Pod
- User provided keychain and user provided policy per fabric interface
- Auto generated keychain and user provided policy per Pod

A node can have multiple policies deployed for more than one fabric link. When this happens, the per fabric interface keychain and policy are given preference on the affected interface. The auto generated keychain and associated MACsec policy are then given the least preference.

APIC MACsec supports two security modes. The MACsec **must secure** only allows encrypted traffic on the link while the **should secure** allows both clear and encrypted traffic on the link. Before deploying MACsec in **must secure** mode, the keychain must be deployed on the affected links or the links will go down. For example, a port can turn on MACsec in **must secure** mode before its peer has received its keychain resulting in the link going down. To address this issue the recommendation is to deploy MACsec in **should secure** mode and once all the links are up then change the security mode to **must secure**.

Note Any MACsec interface configuration change will result in packet drops.

MACsec policy definition consists of configuration specific to keychain definition and configuration related to feature functionality. The keychain definition and feature functionality definitions are placed in separate policies. Enabling MACsec per Pod or per interface involves deploying a combination of a keychain policy and MACsec functionality policy.

Note Using internal generated keychains do not require the user to specify a keychain.

APIC Access MACsec

MACsec encryption is configured under fabric policies for fabric ports and access policies for access ports and can be applied to physical interfaces, port-channels, or vPCs (virtual port-channels). Interfaces with MACsec enabled can operate as either Layer 2 interfaces (such as EPGs or L3Out SVIs) or Layer 3 interfaces (including L3Out routed interfaces or routed sub-interfaces). Enabling MACsec on an interface enables MACsec encryption on the entire interface and applies to all VLANs or sub-interfaces configured on the interface.

Data Plane Policing

Use data plane policing (DPP) to manage bandwidth consumption on ACI fabric access interfaces. DPP policies can apply to egress traffic, ingress traffic, or both. DPP monitors the data rates for a particular interface. When the data rate exceeds user-configured values, marking or dropping of packets occurs immediately. Policing does not buffer the traffic; therefore, the transmission delay is not affected. When traffic exceeds the data rate, the ACI fabric can either drop the packets or mark QoS fields in them.

Note Egress data plane policers are not supported on switched virtual interfaces (SVI).

DPP policies can be single-rate, dual-rate, and color-aware. Single-rate policies monitor the committed information rate (CIR) of traffic. Dual-rate policers monitor both CIR and peak information rate (PIR) of traffic. In addition, the system monitors associated burst sizes. Three colors, or conditions, are determined by the policer for each packet depending on the data rate parameters supplied: conform (green), exceed (yellow), or violate (red).

Typically, DPP policies are applied to physical or virtual layer 2 connections for virtual or physical devices such as servers or hypervisors, and on layer 3 connections for routers. DPP policies applied to leaf switch access ports are configured in the fabric access (`infraInfra`) portion of the ACI fabric, and must be configured by a fabric administrator. DPP policies applied to interfaces on border leaf switch access ports (`l3extOut` or `l2extOut`) are configured in the tenant (`fvTenant`) portion of the ACI fabric, and can be configured by a tenant administrator.

Only one action can be configured for each condition. For example, a DPP policy can to conform to the data rate of 256000 bits per second, with up to 200 millisecond bursts. The system applies the conform action to traffic that falls within this rate, and it would apply the violate action to traffic that exceeds this rate. Color-aware policies assume that traffic has been previously marked with a color. This information is then used in the actions taken by this type of policer.

Scheduler

A schedule allows operations, such as configuration import/export or tech support collection, to occur during one or more specified windows of time.

A schedule contains a set of time windows (occurrences). These windows can be one time only or can recur at a specified time and day each week. The options defined in the window, such as the duration or the maximum number of tasks to be run, determine when a scheduled task executes. For example, if a change cannot be deployed during a given maintenance window because the maximum duration or number of tasks has been reached, that deployment is carried over to the next maintenance window.

Each schedule checks periodically to see whether the APIC has entered one or more maintenance windows. If it has, the schedule executes the deployments that are eligible according to the constraints specified in the maintenance policy.

A schedule contains one or more occurrences, which determine the maintenance windows associated with that schedule. An occurrence can be one of the following:

- One-time Window—Defines a schedule that occurs only once. This window continues until the maximum duration of the window or the maximum number of tasks that can be run in the window has been reached.
- Recurring Window—Defines a repeating schedule. This window continues until the maximum number of tasks or the end of the day specified in the window has been reached.

After a schedule is configured, it can then be selected and applied to the following export and firmware policies during their configuration:

- Tech Support Export Policy
- Configuration Export Policy -- Daily AutoBackup
- Firmware Download

Firmware Upgrade

Policies on the APIC manage the following aspects of the firmware upgrade processes:

- What version of firmware to use.
- Downloading firmware images from Cisco to the APIC repository.
- Compatibility enforcement.
- What to upgrade:
    - Switches
    - The APIC
    - The compatibility catalog
- When the upgrade will be performed.
- How to handle failures (retry, pause, ignore, and so on).

Each firmware image includes a compatibility catalog that identifies supported types and switch models. The APIC maintains a catalog of the firmware images, switch types, and models that are allowed to use that firmware image. The default setting is to reject a firmware update when it does not conform to the compatibility catalog.

The APIC, which performs image management, has an image repository for compatibility catalogs, APIC controller firmware images, and switch images. The administrator can download new firmware images to the APIC image repository from an external HTTP server or SCP server by creating an image source policy.

Firmware Group policies on the APIC define what firmware version is needed.

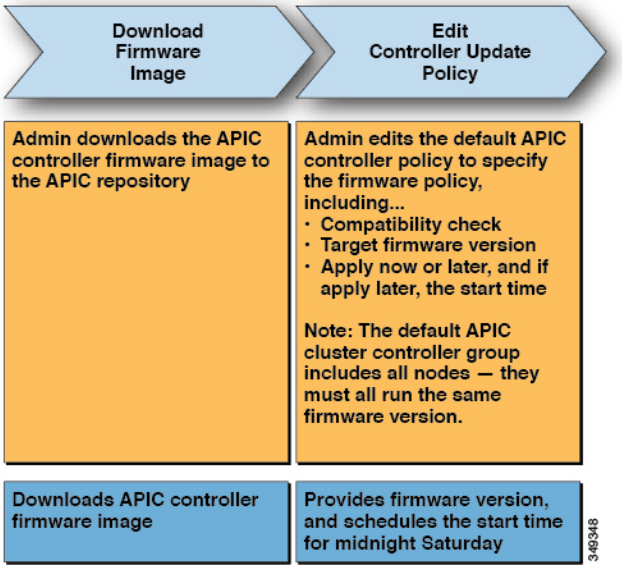
Maintenance Group policies define when to upgrade firmware, which nodes to upgrade, and how to handle failures. In addition, maintenance Group policies define groups of nodes that can be upgraded together and assign those maintenance groups to schedules. Node group options include all leaf nodes, all spine nodes, or sets of nodes that are a portion of the fabric.

The APIC controller firmware upgrade policy always applies to all nodes in the cluster, but the upgrade is always done one node at a time. The APIC GUI provides real-time status information about firmware upgrades.

Note If a recurring or one-time upgrade schedule is set with a date and time in the past, the scheduler triggers the upgrade immediately.

The following figure shows the APIC cluster nodes firmware upgrade process.

Figure 18: APIC Cluster Controller Firmware Upgrade Process



The APIC applies this controller firmware upgrade policy as follows:

- Because the administrator configured the controller update policy with a start time of midnight Saturday, the APIC begins the upgrade at midnight on Saturday.
- The system checks for compatibility of the existing firmware to upgrade to the new version according to the compatibility catalog provided with the new firmware image.
- The upgrade proceeds one node at a time until all nodes in the cluster are upgraded.

Note Because the APIC is a replicated cluster of nodes, disruption should be minimal. An administrator should be aware of the system load when considering scheduling APIC upgrades, and should plan for an upgrade during a maintenance window.

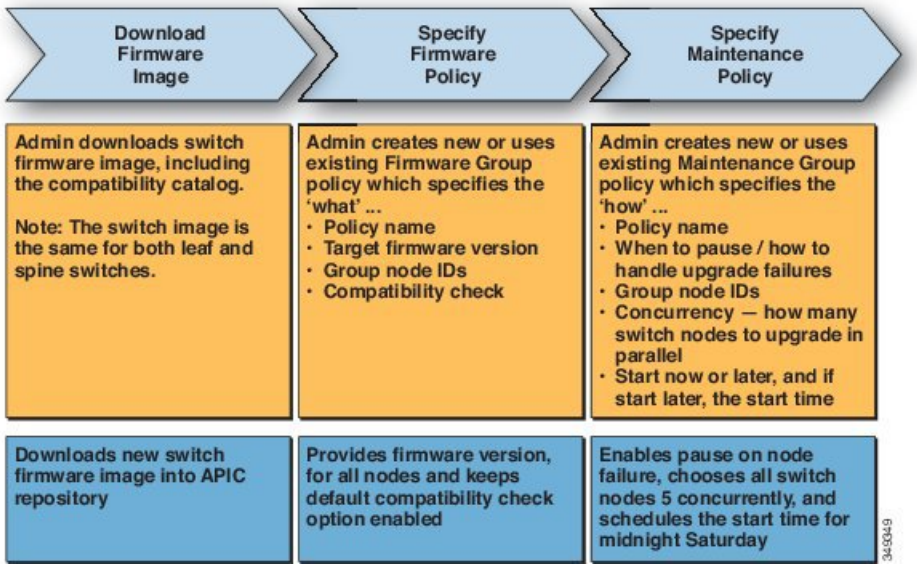
- The ACI fabric, including the APIC, continues to run while the upgrade proceeds.

Note The controllers upgrade in random order. Each APIC controller takes about 10 minutes to upgrade. Once a controller image is upgraded, it drops from the cluster, and it reboots with the newer version while the other APIC controllers in the cluster remain operational. Once the controller reboots, it joins the cluster again. Then the cluster converges, and the next controller image starts to upgrade. If the cluster does not immediately converge and is not fully fit, the upgrade will wait until the cluster converges and is fully fit. During this period, a Waiting for Cluster Convergence message is displayed.

- If a controller node upgrade fails, the upgrade pauses and waits for manual intervention.

The following figure shows how this process works for upgrading all the ACI fabric switch nodes firmware.

Figure 19: Switch Firmware Upgrade Process



The APIC applies this switch upgrade policy as follows:

- Because the administrator configured the controller update policy with a start time of midnight Saturday, the APIC begins the upgrade at midnight on Saturday.
- The system checks for compatibility of the existing firmware to upgrade to the new version according to the compatibility catalog provided with the new firmware image.
- The upgrade proceeds five nodes at a time until all the specified nodes are upgraded.

Note A firmware upgrade causes a switch reboot; the reboot can disrupt the operation of the switch for several minutes. Schedule firmware upgrades during a maintenance window.

- If a switch node fails to upgrade, the upgrade pauses and waits for manual intervention.

Refer to the *Cisco APIC Management, Installation, Upgrade, and Downgrade Guide* for detailed step-by-step instructions for performing firmware upgrades.

Configuration Zones

Configuration zones divide the ACI fabric into different zones that can be updated with configuration changes at different times. This limits the risk of deploying a faulty fabric-wide configuration that might disrupt traffic or even bring the fabric down. An administrator can deploy a configuration to a non-critical zone, and then deploy it to critical zones when satisfied that it is suitable.

The following policies specify configuration zone actions:

- `infrazone:ZoneP` is automatically created upon system upgrade. It cannot be deleted or modified.
- `infrazone:Zone` contains one or more pod groups (`PodGrp`) or one or more node groups (`NodeGrp`).

Note You can only choose `PodGrp` or `NodeGrp`; both cannot be chosen.

A node can be part of only one zone (`infrazone:Zone`). `NodeGrp` has two properties: name, and deployment mode. The deployment mode property can be:

- `enabled` - Pending updates are sent immediately.
- `disabled` - New updates are postponed.

Note

- Do not upgrade, downgrade, commission, or decommission nodes in a disabled configuration zone.
- Do not do a clean reload or an uplink/downlink port conversion reload of nodes in a disabled configuration zone.

- `triggered` - pending updates are sent immediately, and the deployment mode is automatically reset to the value it had before the change to `triggered`.

When a policy on a given set of nodes is created, modified, or deleted, updates are sent to each node where the policy is deployed. Based on policy class and `infrazone` configuration the following happens:.

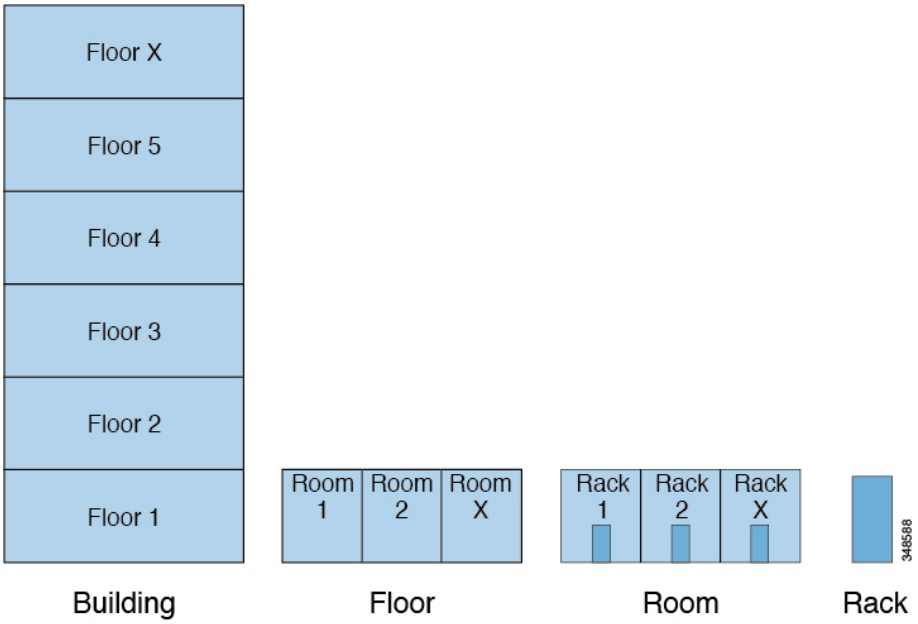
- For policies that do not follow `infrazone` configuration, the APIC sends updates immediately to all the fabric nodes.

- For policies that follow `infrazone` configuration, the update proceeds according to the `infrazone` configuration:
  - If a node is part of an `infrazone:Zone`, the update is sent immediately if the deployment mode of the zone is set to enabled; otherwise the update is postponed.
  - If a node is not part of an`infrazone:Zone`, the update is done immediately, which is the ACI fabric default behavior.

Geolocation

Administrators use geolocation policies to map the physical location of ACI fabric nodes in data center facilities. The following figure shows an example of the geolocation mapping feature.

Figure 20: Geolocation



For example, for fabric deployment in a single room, an administrator would use the default room object, and then create one or more racks to match the physical location of the switches. For a larger deployment, an administrator can create one or more site objects. Each site can contain one or more buildings. Each building has one or more floors. Each floor has one or more rooms, and each room has one or more racks. Finally each rack can be associated with one or more switches.