

Cisco ACI Forwarding

This chapter contains the following sections:

- ACI Fabric Optimizes Modern Data Center Traffic Flows, on page 1
- VXLAN in ACI, on page 2
- Layer 3 VNIDs Facilitate Transporting Inter-subnet Tenant Traffic, on page 4
- Transmission of Spanning Tree Protocol BPDU, on page 6

ACI Fabric Optimizes Modern Data Center Traffic Flows

The Cisco ACI architecture addresses the limitations of traditional data center design, and provides support for the increased east-west traffic demands of modern data centers.

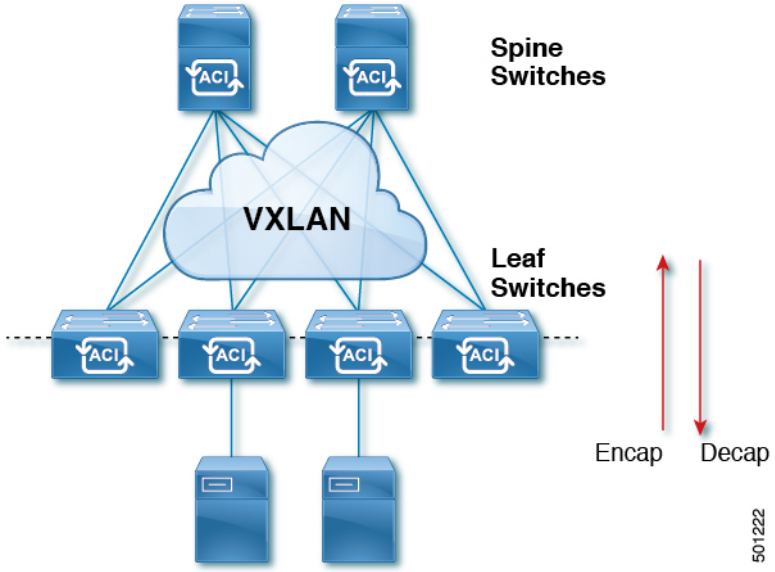
Today, application design drives east-west traffic from server to server through the data center access layer. Applications driving this shift include big data distributed processing designs like Hadoop, live virtual machine or workload migration as with VMware vMotion, server clustering, and multi-tier applications.

North-south traffic drives traditional data center design with core, aggregation, and access layers, or collapsed core and access layers. Client data comes in from the WAN or Internet, a server processes it, and then it exits the data center, which permits data center hardware oversubscription due to WAN or Internet bandwidth constraints. However, Spanning Tree Protocol is required to block loops. This limits available bandwidth due to blocked links, and potentially forces traffic to take a suboptimal path.

In traditional data center designs, IEEE 802.1Q VLANs provide logical segmentation of Layer 2 boundaries or broadcast domains. However, VLAN use of network links is inefficient, requirements for device placements in the data center network can be rigid, and the VLAN maximum of 4094 VLANs can be a limitation. As IT departments and cloud providers build large multi-tenant data centers, VLAN limitations become problematic.

A spine-leaf architecture addresses these limitations. The ACI fabric appears as a single switch to the outside world, capable of bridging and routing. Moving Layer 3 routing to the access layer would limit the Layer 2 reachability that modern applications require. Applications like virtual machine workload mobility and some clustering software require Layer 2 adjacency between source and destination servers. By routing at the access layer, only servers connected to the same access switch with the same VLANs trunked down would be Layer 2-adjacent. In ACI, VXLAN solves this dilemma by decoupling Layer 2 domains from the underlying Layer 3 network infrastructure.

Figure 1: ACI Fabric



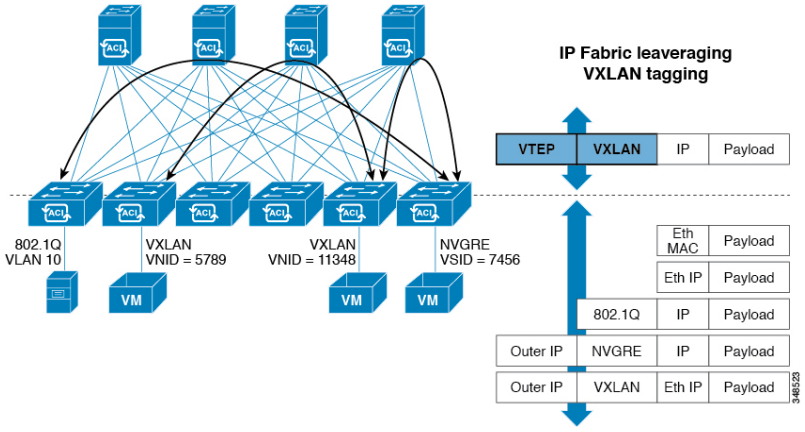
As traffic enters the fabric, ACI encapsulates and applies policy to it, forwards it as needed across the fabric through a spine switch (maximum two-hops), and de-encapsulates it upon exiting the fabric. Within the fabric, ACI uses Intermediate System-to-Intermediate System Protocol (IS-IS) and Council of Oracle Protocol (COOP) for all forwarding of endpoint to endpoint communications. This enables all ACI links to be active, equal cost multipath (ECMP) forwarding in the fabric, and fast-reconverging. For propagating routing information between software defined networks within the fabric and routers external to the fabric, ACI uses the Multiprotocol Border Gateway Protocol (MP-BGP).

VXLAN in ACI

VXLAN is an industry-standard protocol that extends Layer 2 segments over Layer 3 infrastructure to build Layer 2 overlay logical networks. The ACI infrastructure Layer 2 domains reside in the overlay, with isolated broadcast and failure bridge domains. This approach allows the data center network to grow without the risk of creating too large a failure domain.

All traffic in the ACI fabric is normalized as VXLAN packets. At ingress, ACI encapsulates external VLAN, VXLAN, and NVGRE packets in a VXLAN packet. The following figure shows ACI encapsulation normalization.

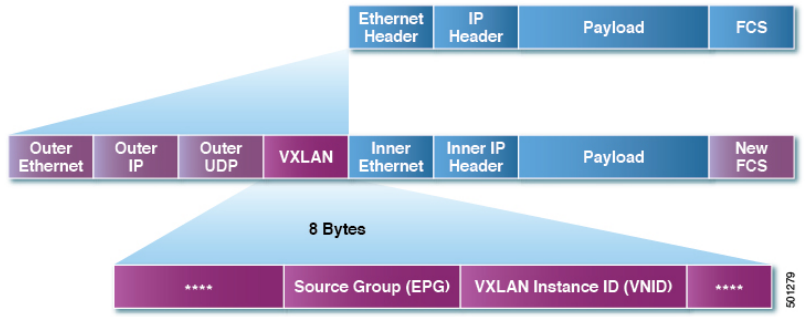
Figure 2: ACI Encapsulation Normalization



Forwarding in the ACI fabric is not limited to or constrained by the encapsulation type or encapsulation overlay network. An ACI bridge domain forwarding policy can be defined to provide standard VLAN behavior where required.

Because every packet in the fabric carries ACI policy attributes, ACI can consistently enforce policy in a fully distributed manner. ACI decouples application policy EPG identity from forwarding. The following illustration shows how the ACI VXLAN header identifies application policy within the fabric.

Figure 3: ACI VXLAN Packet Format



The ACI VXLAN packet contains both Layer 2 MAC address and Layer 3 IP address source and destination fields, which enables efficient and scalable forwarding within the fabric. The ACI VXLAN packet header source group field identifies the application policy endpoint group (EPG) to which the packet belongs. The VXLAN Instance ID (VNID) enables forwarding of the packet through tenant virtual routing and forwarding (VRF) domains within the fabric. The 24-bit VNID field in the VXLAN header provides an expanded address space for up to 16 million unique Layer 2 segments in the same network. This expanded address space gives IT departments and cloud providers greater flexibility as they build large multitenant data centers.

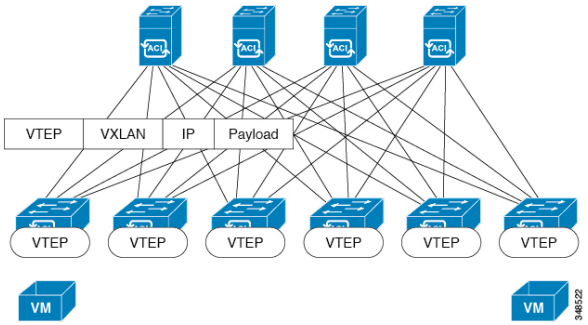
VXLAN enables ACI to deploy Layer 2 virtual networks at scale across the fabric underlay Layer 3 infrastructure. Application endpoint hosts can be flexibly placed in the data center network without concern for the Layer 3 boundary of the underlay infrastructure, while maintaining Layer 2 adjacency in a VXLAN overlay network.

Layer 3 VNIDs Facilitate Transporting Inter-subnet Tenant Traffic

The ACI fabric provides tenant default gateway functionality that routes between the ACI fabric VXLAN networks. For each tenant, the fabric provides a virtual default gateway that spans all of the leaf switches assigned to the tenant. It does this at the ingress interface of the first leaf switch connected to the endpoint. Each ingress interface supports the default gateway interface. All of the ingress interfaces across the fabric share the same router IP address and MAC address for a given tenant subnet.

The ACI fabric decouples the tenant endpoint address, its identifier, from the location of the endpoint that is defined by its locator or VXLAN tunnel endpoint (VTEP) address. Forwarding within the fabric is between VTEPs. The following figure shows decoupled identity and location in ACI.

Figure 4: ACI Decouples Identity and Location



VXLAN uses VTEP devices to map tenant end devices to VXLAN segments and to perform VXLAN encapsulation and de-encapsulation. Each VTEP function has two interfaces:

- A switch interface on the local LAN segment to support local endpoint communication through bridging
- An IP interface to the transport IP network

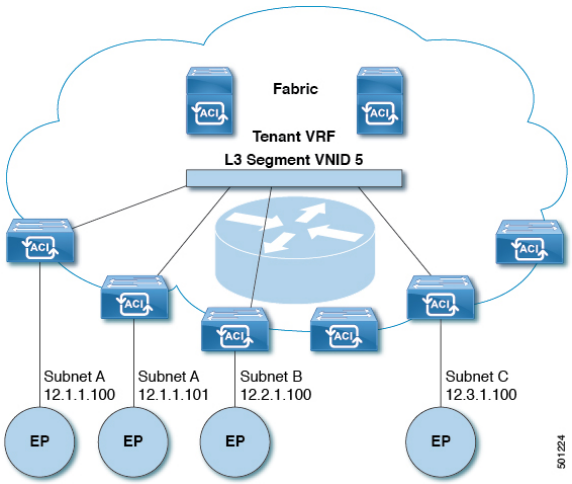
The IP interface has a unique IP address that identifies the VTEP device on the transport IP network known as the infrastructure VLAN. The VTEP device uses this IP address to encapsulate Ethernet frames and transmit the encapsulated packets to the transport network through the IP interface. A VTEP device also discovers the remote VTEPs for its VXLAN segments and learns remote MAC Address-to-VTEP mappings through its IP interface.

The VTEP in ACI maps the internal tenant MAC or IP address to a location using a distributed mapping database. After the VTEP completes a lookup, the VTEP sends the original data packet encapsulated in VXLAN with the destination address of the VTEP on the destination leaf switch. The destination leaf switch de-encapsulates the packet and sends it to the receiving host. With this model, ACI uses a full mesh, single hop, loop-free topology without the need to use the spanning-tree protocol to prevent loops.

The VXLAN segments are independent of the underlying network topology; conversely, the underlying IP network between VTEPs is independent of the VXLAN overlay. It routes the encapsulated packets based on the outer IP address header, which has the initiating VTEP as the source IP address and the terminating VTEP as the destination IP address.

The following figure shows how routing within the tenant is done.

Figure 5: Layer 3 VNIDs Transport ACI Inter-subnet Tenant Traffic



For each tenant VRF in the fabric, ACI assigns a single L3 VNID. ACI transports traffic across the fabric according to the L3 VNID. At the egress leaf switch, ACI routes the packet from the L3 VNID to the VNID of the egress subnet.

Traffic arriving at the fabric ingress that is sent to the ACI fabric default gateway is routed into the Layer 3 VNID. This provides very efficient forwarding in the fabric for traffic routed within the tenant. For example, with this model, traffic between 2 VMs belonging to the same tenant, on the same physical host, but on different subnets, only needs to travel to the ingress switch interface before being routed (using the minimal path cost) to the correct destination.

To distribute external routes within the fabric, ACI route reflectors use multiprotocol BGP (MP-BGP). The fabric administrator provides the autonomous system (AS) number and specifies the spine switches that become route reflectors.

Note Cisco ACI does not support IP fragmentation. Therefore, when you configure Layer 3 Outside (L3Out) connections to external routers, or Multi-Pod connections through an Inter-Pod Network (IPN), it is recommended that the interface MTU is set appropriately on both ends of a link.

IGP Protocol Packets (EIGRP, OSPFv3) are constructed by components based on the Interface MTU size. In Cisco ACI, if the CPU MTU size is less than the Interface MTU size and if the constructed packet size is greater than the CPU MTU, then the packet is dropped by the kernal, especially in IPv6. To avoid such control packet drops always configure the same MTU values on both the control plane and on the interface.

On some platforms, such as Cisco ACI, Cisco NX-OS, and Cisco IOS, the configurable MTU value does not take into account the Ethernet headers (matching IP MTU, and excluding the 14-18 Ethernet header size), while other platforms, such as IOS-XR, include the Ethernet header in the configured MTU value. A configured value of 9000 results in a max IP packet size of 9000 bytes in Cisco ACI, Cisco NX-OS, and Cisco IOS, but results in a max IP packet size of 8986 bytes for an IOS-XR untagged interface.

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

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

Transmission of Spanning Tree Protocol BPDU

When there are two or more switches running the spanning tree protocol (STP) that are connected to Cisco Application Centric Infrastructure (ACI) in an EPG and if the static port assignments are as follows:

- All statically assigned ports under EPG are access untagged: STP bridge protocol data units (BPDUs) are received and sent without a tag.
- For a mix of statically assigned trunk ports and statically assigned access untagged ports: STP BPDUs received on a trunk port will be sent to the access untagged port with the `dot1q` tag. Hence the access port will go into a inconsistent state.
- For a mix of statically assigned trunk ports and statically assigned access ports under an EPG: Cisco ACI sends STP BPDU with `dot1q` tag and the access ports use `802.1p` access.

  In this case you must use `802.1p` access on the Layer 2 switch to receive and process the tagged STP packets.

  In case `802.1p` is not allowed on the Layer 2 switch, then use the trunk port access.
- Cisco ACI acts as a full-duplex hub and will flood spanning tree BPDUs within the VxLAN VNID associated to the encapsulation VLAN on which the BPDUs were received. Because Cisco ACI is a full-duplex medium, external switches running versions of rapid spanning tree protocol (RSTP) or rapid per VLAN spanning tree (RPVST) will default to point-to-point link types. As a result, if there are more than 2 external switches that run STP and connect to the same encapsulation VLAN and EPG VNID, you should set the link type to "shared" on the external switch interfaces to avoid problems with convergence and instability. These problems can occur because the switches will receive BPDUs from all bridges (or STP-enabled switches) that are connected to this encapsulation.

Spanning tree BPDUs are flooded within a given VLAN ID defined on an EPG path. This VLAN is referred to as `FD_VLAN` on leaf switches. To forward traffic within the FD_VLAN across leaf switches, Cisco ACI allocates a VXLAN VNID called the `fabric_encap`. The `fabric_encap` is derived by taking a numeric base identifier that belongs to a VLAN pool and then adding the index value for the VLAN ID that was allocated from the VLAN pool. For example, VXLAN VNID 9000 is assigned to the VLAN pool A that contains VLAN range 10-20. VLAN 10 from VLAN pool A is assigned VNID 9000, VLAN 11 is assigned VNID 9001.

For this reason, if two separate EPGs are using the same VLAN ID and are allocating that VLAN ID from the same VLAN pool, it is possible that the same `fabric_encap` VNID can be derived for two EPGs on different leaf switches. This creates the potential for spanning tree BPDUs to be unintentionally flooded between the two EPGs.

This behavior can be avoided by ensuring that different domains, such as physical domain, with separate VLAN pools are assigned to each EPG so that it allocates a given VLAN ID from separate VLAN pools. This ensures that the base identifier will be different and prevent overlapping `fabric_encap` VNIDs from being allocated.

The `fabric_encap` values can be checked using the following command and checking the "Fabric_enc" column in the output on leaf switches for a given 802.1q VLAN ID:

```
vsh_lc -c "show system internal eltmc info vlan br"
```