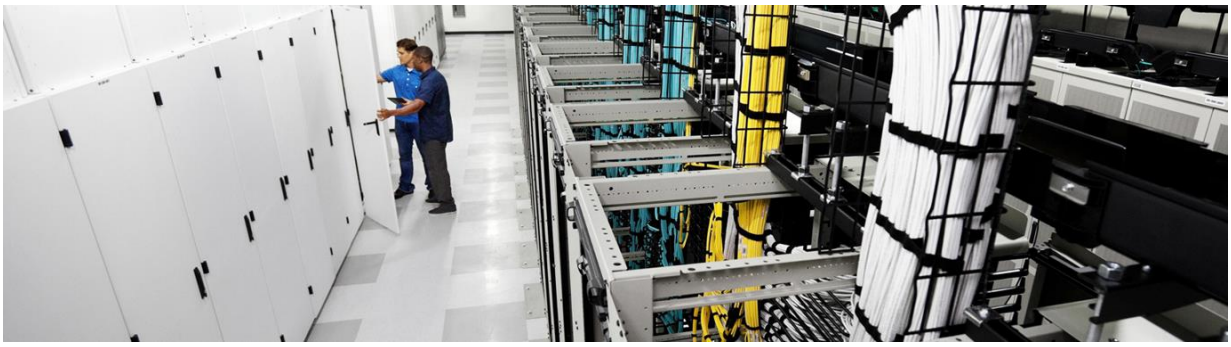
White Paper

VXLAN Overview: Cisco Nexus 9000 Series Switches



What You Will Learn

Traditional network segmentation has been provided by VLANs that are standardized under the IEEE 802.1Q group. VLANs provide logical segmentation of Layer 2 boundaries or broadcast domains. However, due to the inefficient use of available network links with VLAN use, rigid requirements on device placements in the data center network, and the limited scalability to a maximum 4094 VLANs, using VLANs has become a limiting factor to IT departments and cloud providers as they build large multitenant data centers.

Cisco, in partnership with other leading vendors, proposed the Virtual Extensible LAN (VXLAN) standard to the IETF as a solution to the data center network challenges posed by traditional VLAN technology. The VXLAN standard provides for the elastic workload placement and higher scalability of Layer2 segmentation that is required by today's application demands.

The Cisco Nexus® 9000 Series of switches consists of Cisco Nexus 9500 platform modular switches and Cisco Nexus 9300 platform fixed-configuration switches. They are designed for the next-generation data center with industry-leading hardware-based VXLAN function, which provides Layer 2 connectivity extension across the Layer 3 boundary and easy integration between VXLAN and non-VXLAN infrastructures. They enable large-scale virtualized and multitenant data center designs over a shared common physical infrastructure.

Cisco Nexus 9000 Series Switches can run in ACI mode or NX-OS mode. In ACI mode Cisco Nexus 9000 Series Switches when used in combination with the Cisco Application Policy Infrastructure Controller (APIC) provide an application centric infrastructure. In NX-OS mode Cisco Nexus 9000 Series Switches function as classic switches. Equipped with enhanced Cisco® NX-OS Software as the operating system, Cisco Nexus 9000 Series Switches provide network connectivity through traditional means but with exceptional performance and enhanced network resiliency and programmatic automation functions. The VXLAN implementations on Cisco Nexus 9000 Series Switches differ between ACI mode and NX-OS mode. This white paper provides an overview of VXLAN on Cisco Nexus 9000 Series Switches in NX-OS mode.

© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

VXLAN Overview

As its name indicates, VXLAN is designed to provide the same Ethernet Layer 2 network services as VLAN does today, but with greater extensibility and flexibility. Compared to VLAN, VXLAN offers the following benefits:

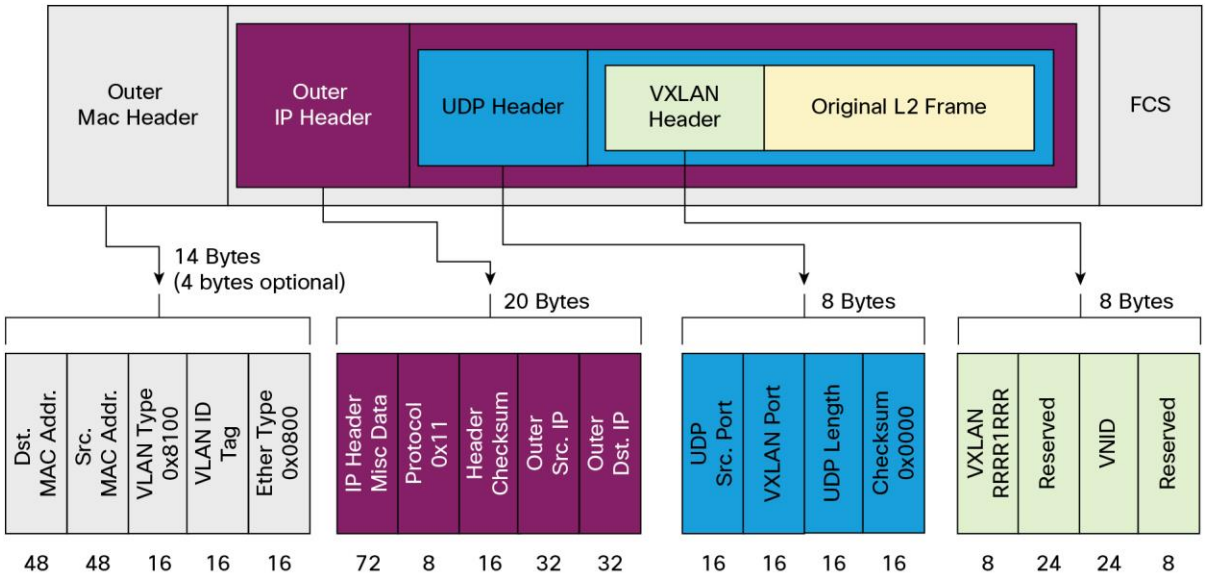
- Flexible placement of multitenant segments throughout the data center: It provides a solution to extend Layer 2 segments over the underlying shared network infrastructure so that tenant workload can be placed across physical pods in the data center.
- Higher scalability to address more Layer 2 segments: VLANs use a 12-bit VLAN ID to address Layer 2 segments, which results in limiting scalability of only 4094 VLANs. VXLAN uses a 24-bit segment ID known as the VXLAN network identifier (VNID), which enables up to 16 million VXLAN segments to coexist in the same administrative domain.
- Better utilization of available network paths in the underlying infrastructure: VLAN uses the Spanning Tree Protocol for loop prevention, which ends up not using half of the network links in a network by blocking redundant paths. In contrast, VXLAN packets are transferred through the underlying network based on its Layer 3 header and can take complete advantage of Layer 3 routing, equal-cost multipath (ECMP) routing, and link aggregation protocols to use all available paths.

VXLAN Encapsulation and Packet Format

VXLAN is a Layer 2 overlay scheme over a Layer 3 network. It uses MAC Address-in-User Datagram Protocol (MAC-in-UDP) encapsulation to provide a means to extend Layer 2 segments across the data center network. VXLAN is a solution to support a flexible, large-scale multitenant environment over a shared common physical infrastructure. The transport protocol over the physical data center network is IP plus UDP.

VXLAN defines a MAC-in-UDP encapsulation scheme where the original Layer 2 frame has a VXLAN header added and is then placed in a UDP-IP packet. With this MAC-in-UDP encapsulation, VXLAN tunnels Layer 2 network over Layer 3 network. The VXLAN packet format is shown in Figure 1.

Figure 1. VXLAN Packet Format



© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

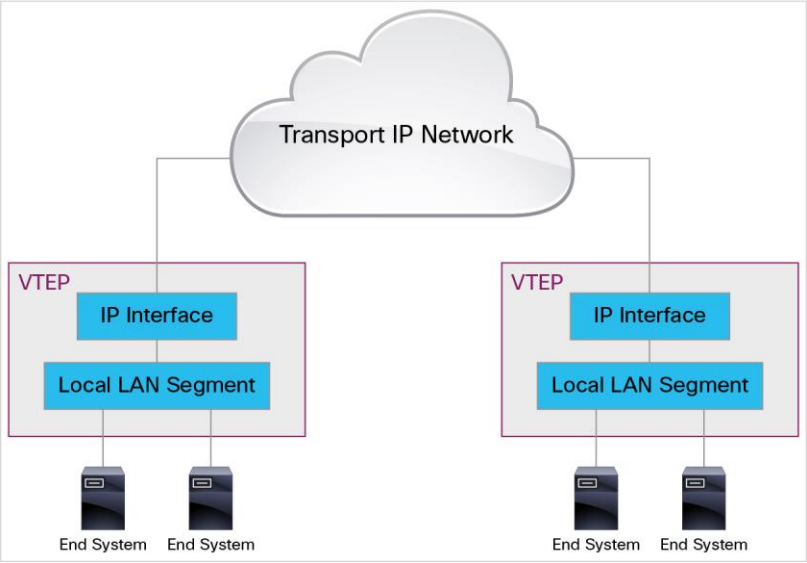
As shown in Figure 1, VXLAN introduces an 8-byte VXLAN header that consists of a 24-bit VNID and a few reserved bits. The VXLAN header together with the original Ethernet frame goes in the UDP payload. The 24-bit VNID is used to identify Layer 2 segments and to maintain Layer 2 isolation between the segments. With all 24 bits in VNID, VXLAN can support 16 million LAN segments.

VXLAN Tunnel Endpoint

VXLAN uses VXLAN tunnel endpoint (VTEP) devices to map tenants' end devices to VXLAN segments and to perform VXLAN encapsulation and de-encapsulation. Each VTEP function has two interfaces: One is a switch interface on the local LAN segment to support local endpoint communication through bridging, and the other is 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 transmits 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 functional components of VTEPs and the logical topology that is created for Layer 2 connectivity across the transport IP network is shown in Figure 2.

Figure 2. VTEP



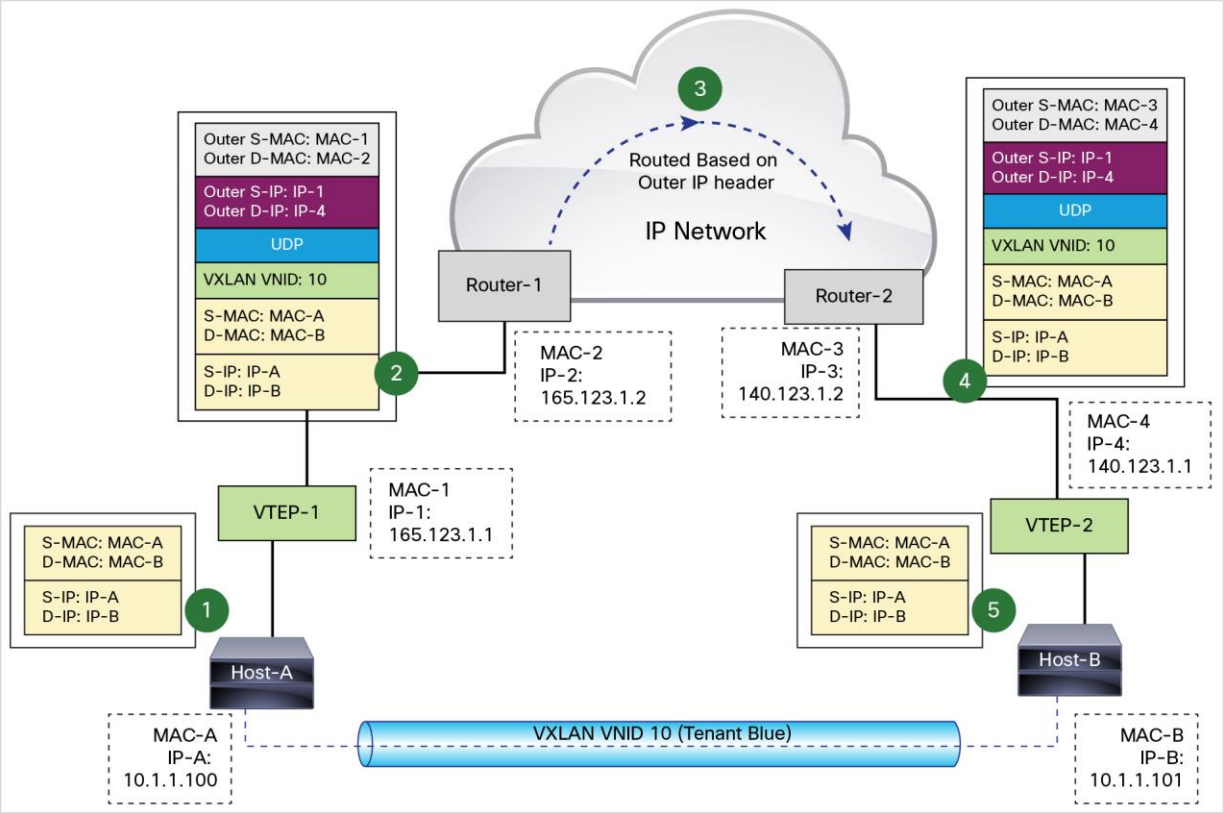
The VXLAN segments are independent of the underlying network topology; conversely, the underlying IP network between VTEPs is independent of the VXLAN overlay. It routes the encapsulated packets based on the outer IP address header, which has the initiating VTEP as the source IP address and the terminating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

© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

VXLAN Packet Forwarding Flow

VXLAN uses stateless tunnels between VTEPs to transmit traffic of the overlay Layer 2 network through the Layer 3 transport network. An example of a VXLAN packet forwarding flow is shown in Figure 3.

Figure 3. VXLAN Unicast Packet Forwarding Flow



In Figure 3, Host-A and Host-B in VXLAN segment 10 communicate with each other through the VXLAN tunnel between VTEP-1 and VTEP-2. This example assumes that address learning has been done on both sides, and corresponding MAC-to-VTEP mappings exist on both VTEPs.

When Host-A sends traffic to Host-B, it forms Ethernet frames with MAC-B address of Host-B as the destination MAC address and sends them out to VTEP-1. VTEP-1, with a mapping of MAC-B to VTEP-2 in its mapping table, performs VXLAN encapsulation on the packets by adding VXLAN, UDP, and outer IP address header to it. In the outer IP address header, the source IP address is the IP address of VTEP-1, and the destination IP address is the IP address of VTEP-2. VTEP-1 then performs an IP address lookup for the IP address of VTEP-2 to resolve the next hop in the transit network and subsequently uses the MAC address of the next-hop device to further encapsulate the packets in an Ethernet frame to send to the next-hop device.

The packets are routed toward VTEP-2 through the transport network based on their outer IP address header, which has the IP address of VTEP-2 as the destination address. After VTEP-2 receives the packets, it strips off the outer Ethernet, IP, UDP, and VXLAN headers, and forwards the packets to Host-B, based on the original destination MAC address in the Ethernet frame.

© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

VXLAN Implementation on Cisco Nexus 9000 Series Switches

Cisco Nexus 9000 Series Switches support the hardware-based VXLAN function that extends Layer 2 connectivity across the Layer 3 transport network and provides a high-performance gateway between VXLAN and non-VXLAN infrastructures. The following sections provide the details of VXLAN implementation on Cisco Nexus 9000 Series Switches in NX-OS mode.

Layer 2 Mechanisms for Broadcast, Unknown Unicast, and Multicast Traffic

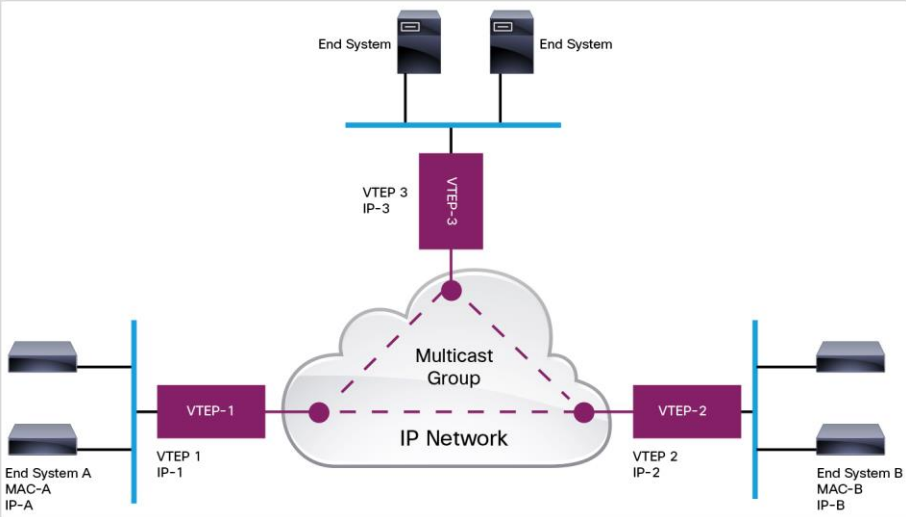
VXLAN on Cisco Nexus 9000 Series Switches uses existing Layer 2 mechanisms - flooding, and dynamic MAC address learning - to do the following:

- Transport broadcast, unknown unicast, and multicast traffic
- Discover remote VTEPs
- Learn remote host MAC addresses and MAC-to-VTEP mappings for each VXLAN segment

For these traffic types, IP multicast is used to reduce the flooding scope of the set of hosts that are participating in the VXLAN segment.

Each VXLAN segment, or VNID, is mapped to an IP multicast group in the transport IP network. Each VTEP device is independently configured and joins this multicast group as an IP host through the Internet Group Management Protocol (IGMP). The IGMP joins trigger Protocol Independent Multicast (PIM) joins and signaling through the transport network for the particular multicast group. The multicast distribution tree for this group is built through the transport network based on the locations of participating VTEPs. The multicast tunnel of a VXLAN segment through the underlying IP network is shown in Figure 4.

Figure 4. VXLAN Multicast Group in Transport Network



This multicast group shown in Figure 4 is used to transmit VXLAN broadcast, unknown unicast, and multicast traffic through the IP network, limiting Layer 2 flooding to those devices that have end systems participating in the same VXLAN segment. VTEPs communicate with one another through the flooded or multicast traffic in this multicast group.

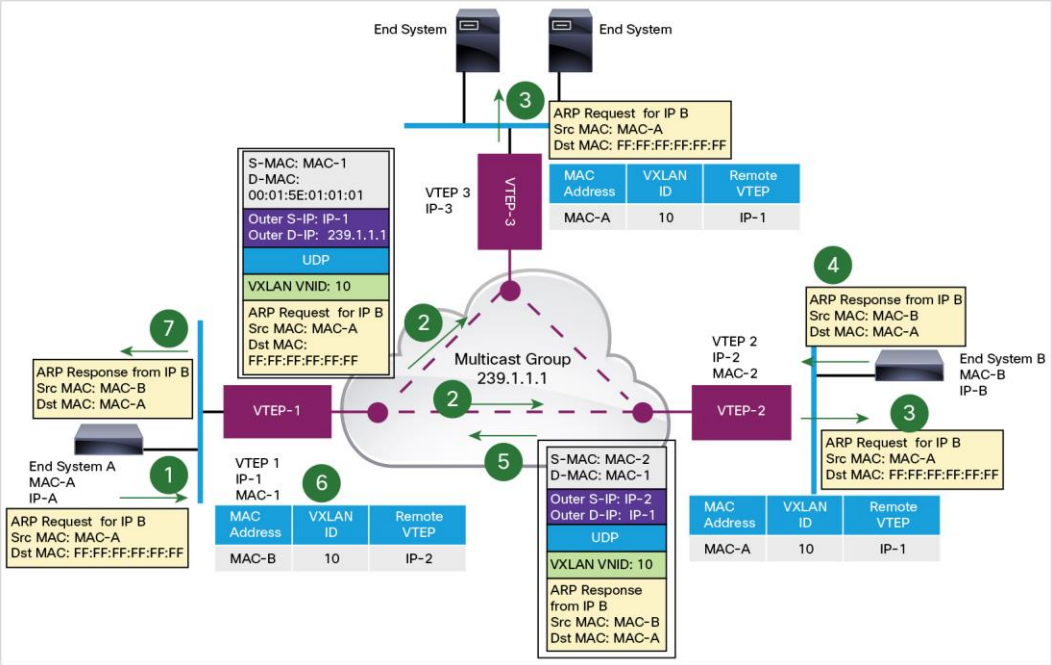
© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Remote VTEP Discovery and Tenant Address Learning

The Cisco Nexus 9000 VXLAN implementation uses the classic Layer 2 data plane flooding and learning mechanisms for remote VTEP discovery and tenant address learning. The network in Figure 4 is an example that shows the learning process.

The tenant VXLAN segment has VNID 10 and uses the multicast group 239.1.1.1 over the transport network. It has three participating VTEPs in the data center. Assume that no address learning has been performed between locations. End System A (with IP-A, MAC-A) starts IP communication with End System B (with IP-B, MAC-B). The sequence of steps is shown in Figure 5.

Figure 5. VXLAN Peer Discoveries and Tenant Address Learning



1. End System A sends out an Address Resolution Protocol (ARP) request for IP-B on its Layer 2 VXLAN network.
2. VTEP-1 receives the ARP request. It does not yet have a mapping for IP-B. VTEP-1 encapsulates the ARP request in an IP multicast packet and forwards it to the VXLAN multicast group. The encapsulated multicast packet has the IP address of VTEP-1 as the source IP address and the VXLAN multicast group address as the destination IP address.
3. The IP multicast packet is distributed to all members in the tree. VTEP-2 and VTEP-3 receive the encapsulated multicast packet because they've joined the VXLAN multicast group.

   They de-encapsulate the packet and check its VNID in the VXLAN header. If it matches their configured VXLAN segment VNID, they forward the ARP request to their local VXLAN network. They also learn the IP address of VTEP-1 from the outer IP address header and inspect the packet to learn the MAC address of End System A, placing this mapping in the local table.
4. End System B receives the ARP request forwarded by VTEP-2. It responds with its own MAC address (MAC-B), and learns the IP-A-to-MAC-A mapping.

© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

5. VTEP-2 receives the ARP reply of End System B that has MAC-A as the destination MAC address. It now knows about MAC-A-to-IP-1 mapping. It can use the unicast tunnel to forward the ARP reply back to VTEP-1. In the encapsulated unicast packet, the source IP address is IP-2 and the destination IP address is IP-1. The ARP reply is encapsulated in the UDP payload.
6. VTEP-1 receives the encapsulated ARP reply from VTEP-2. It de-encapsulates and forwards the ARP reply to End System A. It also learns the IP address of VTEP-2 from the outer IP address header and inspects the original packet to learn MAC-B-to-IP-2 mapping.
7. Subsequent IP packets between End Systems A and B are unicast forwarded, based on the mapping information on VTEP-1 and VTEP-2, using the VXLAN tunnel between them.
8. VTEP-1 can optionally perform proxy ARPs for subsequent ARP requests for IP-B to reduce the flooding over the transport network.

ECMP and LACP Load Sharing with VXLAN

Encapsulated VXLAN packets are forwarded between VTEPs based on the native forwarding decisions of the transport network. Most of the data center transport networks are designed and deployed with multiple redundant paths and take advantage of various multipath load-sharing technologies to distribute traffic loads on all available paths. It is desirable to share the load of the VXLAN traffic in the same fashion in the transport network.

A typical VXLAN transport network is an IP-routing network that uses the standard IP ECMP to balance the traffic load among multiple best paths. To avoid out-of-sequence packet forwarding, flow-based ECMP is commonly deployed. An ECMP flow is defined by the source and destination IP addresses and optionally the source and destination TCP or UDP ports in the IP packet header.

Because all the VXLAN packet flows between a pair of VTEPs have the same outer source and destination IP addresses, and all VTEP devices must use one identical destination UDP port that can be either the Internet Allocated Numbers Authority (IANA)-allocated UDP port 4789 or a customer-configured port, the only variable element in the ECMP flow definition that can differentiate VXLAN flows from the transport network standpoint is the source UDP port. A similar situation for Link Aggregation Control Protocol (LACP) hashing occurs if the resolved egress interface based on the routing and ECMP decision is an LACP port channel. The LACP uses the VXLAN outer-packet header for link load-share hashing, which results in the source UDP port being the only element that can uniquely identify a VXLAN flow.

Cisco Nexus 9000 Series Switches implement VXLAN in the way that a hash of the inner frame's header is used as the VXLAN source UDP port. As a result, a VXLAN flow can be unique, with the IP addresses and UDP ports combination in its outer header while traversing the underlay transport network. Therefore, the hashed source UDP port introduces a desirable level of entropy for ECMP and LACP load balancing.

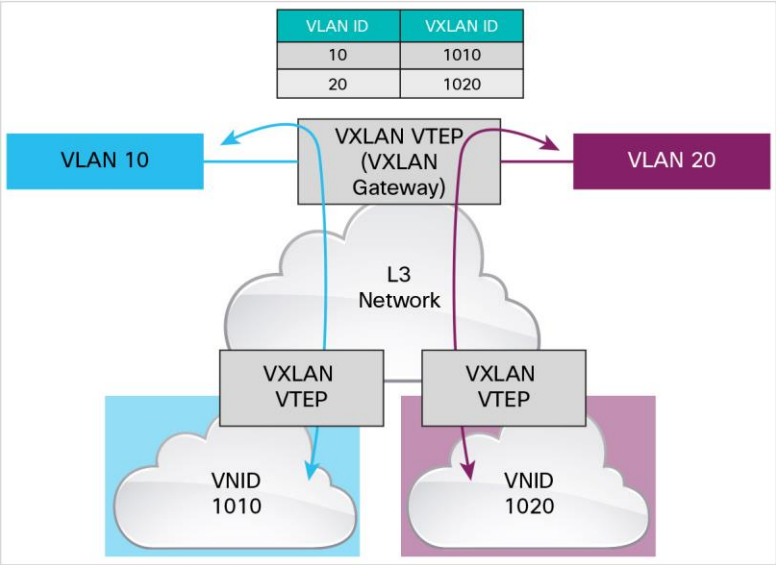
Cisco Nexus 9000 as Hardware-Based VXLAN Gateway

VXLAN is a new technology for virtual data center overlays and is being adopted in data center networks more and more, especially for virtual networking in the hypervisor for virtual machine-to-virtual machine communication. However, data centers are likely to contain devices that are not capable of supporting VXLAN, such as legacy hypervisors, physical servers, and network services appliances, such as physical firewalls and load balancers, and storage devices, etc. Those devices need to continue to reside on classic VLAN segments. It is not uncommon that virtual machines in a VXLAN segment need to access services provided by devices in a classic VLAN segment. This type of VXLAN-to-VLAN connectivity is enabled by using a VXLAN gateway.

© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

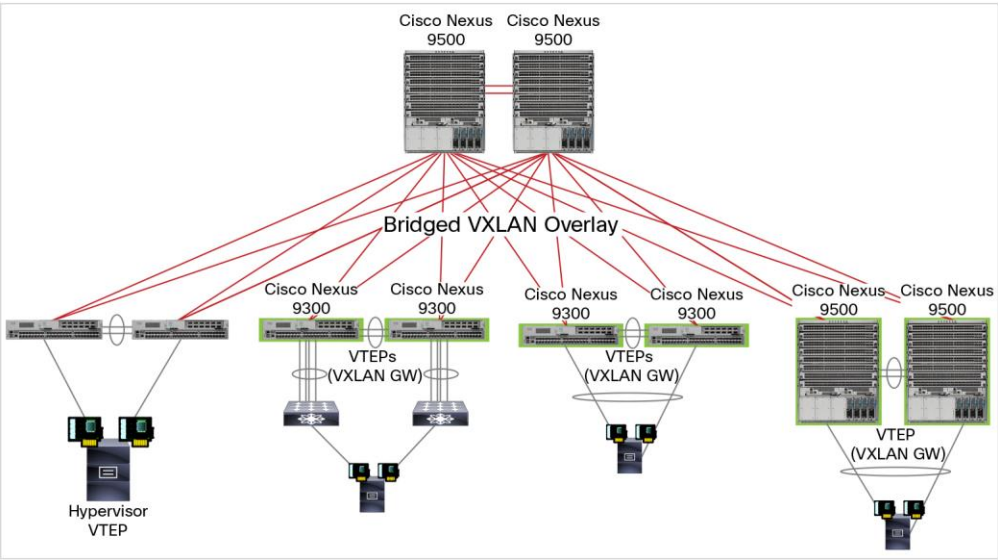
A VXLAN gateway is a VTEP device that combines a VXLAN segment and a classic VLAN segment into one common Layer 2 domain. The logic mapping between IEEE 802.1Q VLAN and VXLAN on a VXLAN gateway is shown in Figure 6.

Figure 6. VXLAN-to-VLAN Logic Mapping by VXLAN Gateway



A Cisco Nexus 9000 Series Switch can function as a hardware-based VXLAN gateway. It seamlessly connects VXLAN and VLAN segments as one forwarding domain across the Layer 3 boundary without sacrificing forwarding performance. The Cisco Nexus 9000 Series eliminates the need for an additional physical or virtual device to be the gateway. The hardware-based encapsulation and de-encapsulation provides line-rate performance for all frame sizes. Examples of Cisco Nexus 9000 Series Switches as VXLAN gateways are shown in Figure 7.

Figure 7. Cisco Nexus 9000 as VXLAN Gateways



© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Network Considerations for Common VXLAN Deployments

MTU Size in the Transport Network

Due to the MAC-to-UDP encapsulation, VXLAN introduces 50-byte overhead to the original frames. Therefore, the maximum transmission unit (MTU) in the transport network needs to be increased by 50 bytes. If the overlays use a 1500-byte MTU, the transport network needs to be configured to accommodate 1550-byte packets at a minimum. Jumbo-frame support in the transport network is required if the overlay applications tend to use larger frame sizes than 1500 bytes.

ECMP and LACP Hashing Algorithms in the Transport Network

As described in a previous section, Cisco Nexus 9000 Series Switches introduce a level of entropy in the source UDP port for ECMP and LACP hashing in the transport network. As a way to augment this implementation, the transport network uses an ECMP or LACP hashing algorithm that takes the UDP source port as an input for hashing, which achieves the best load-sharing results for VXLAN encapsulated traffic.

Multicast Group Scaling

The VXLAN implementation on Cisco Nexus 9000 Series Switches uses multicast tunnels for broadcast, unknown unicast, and multicast traffic forwarding. Ideally, one VXLAN segment mapping to one IP multicast group is the way to provide the optimal multicast forwarding. It is possible, however, to have multiple VXLAN segments share a single IP multicast group in the core network.

VXLAN can support up to 16 million logical Layer 2 segments, using the 24-bit VNID field in the header. With one-to-one mapping between VXLAN segments and IP multicast groups, an increase in the number of VXLAN segments causes a parallel increase in the required multicast address space and the amount of forwarding states on the core network devices. At some point, multicast scalability in the transport network can become a concern. In this case, mapping multiple VXLAN segments to a single multicast group can help conserve multicast control plane resources on the core devices and achieve the desired VXLAN scalability. However, this mapping comes at the cost of suboptimal multicast forwarding. Packets forwarded to the multicast group for one tenant are now sent to the VTEPs of other tenants that are sharing the same multicast group. This causes inefficient utilization of multicast data plane resources. Therefore, this solution is a trade-off between control plane scalability and data plane efficiency.

Despite the suboptimal multicast replication and forwarding, having multiple-tenant VXLAN networks to share a multicast group does not bring any implications to the Layer 2 isolation between the tenant networks. After receiving an encapsulated packet from the multicast group, a VTEP checks and validates the VNID in the VXLAN header of the packet. The VTEP discards the packet if the VNID is unknown to it. Only when the VNID matches one of the VTEP's local VXLAN VNIDs, does it forward the packet to that VXLAN segment. Other tenant networks will not receive the packet. Thus, the segregation between VXLAN segments is not compromised.

Conclusion

VXLAN provides a solution to extend Layer 2 networks across Layer 3 infrastructure by way of MAC-in-UDP encapsulation and tunneling. VXLAN enables flexible workload placements by way of the Layer 2 extension. It is also an approach to building a multitenant data center by decoupling tenant Layer 2 segments from the shared transport network.

© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Virtualized hosts are increasingly adopting VXLAN; however, it is rare to have a completely virtualized environment in a data center. More commonly, a data center has coexisting virtual machines, bare-metal hosts, and physical service appliances. Virtual machines need to access services on physical hosts and appliances, which creates the need of a gateway for virtual machines in a VXLAN segment to communicate with devices in a classic VLAN segment.

Cisco Nexus 9000 Series Switches support VXLAN functions with hardware-based performance. Deployed as a VXLAN gateway, Cisco Nexus 9000 Series Switches easily connect VXLAN and classic VLAN segments to create a common forwarding domain so that tenant devices can flexibly reside in both environments with virtually one-hop connectivity. In contrast to software-based VXLAN gateway solutions, Cisco Nexus 9000 Series Switches provide line-rate performance in hardware, which is critical to ensuring the performance of applications that involves devices in both VXLAN and VLAN networks.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

Printed in USA C11-729383-01 01/15

© 2015 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.