White Paper

Cisco Border Gateway Protocol Control Plane for Virtual Extensible LAN

Scalable, Standards-Based Solution for Multitenant Data Centers

What You Will Learn

The Cisco® Border Gateway Protocol (BGP) Control Plane for Virtual Extensible LAN (VXLAN) is an innovative solution that uses Cisco's proven track record for delivering scalable BGP solutions for multitenant data centers. By coupling this standards-based control plane with the Cisco Nexus® 9000 Series Switches, which uniquely support VXLAN Layer 3 forwarding, we provide our customers with the flexibility to incrementally build a scalable multitenant data center based on their needs.

Cloud Computing and Multitenancy

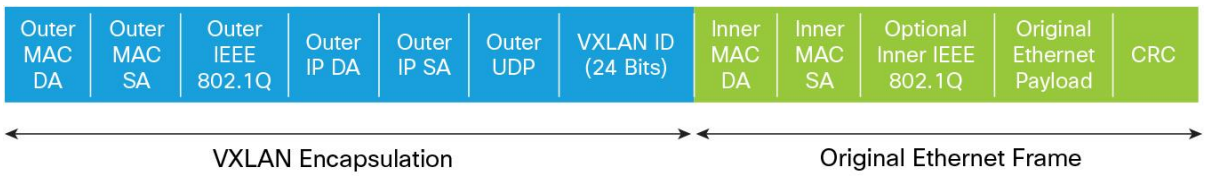
Cloud computing has gained momentum and broad acceptance for its compelling economic advantages in both revenue and profitability. But among the many needs for a cloud infrastructure are the critical multitenancy requirements of network segmentation, traffic separation, elasticity, and workload mobility.

VXLAN is an overlay technology that provides Layer 2 connectivity for workloads residing at noncontiguous points in the data center network. It overcomes the 4092-segment limitation of VLANs and allows for an infrastructure that can scale to 16 million tenants. In addition, VXLAN enables flexibility by allowing workloads to be placed anywhere, along with the traffic separation required in a multitenant environment. However, the VXLAN IETF draft does not specify a control plane, and relies on a flood-and-learn mechanism for host and endpoint discovery. The Cisco BGP Control Plane for VXLAN solution uses the proven features of BGP to provide a more scalable, flexible, and policy-based alternative.

Overview of VXLAN

VXLAN is a MAC address-in-User Datagram Protocol (UDP) tunneling mechanism that identifies the Layer 2 segment through a 24-bit segment identifier called the VXLAN network identifier (VNI) (Figure 1). The larger VNI range allows the LAN to scale to 16 million segments in a cloud network. In addition, IP and UDP encapsulation allows each LAN segment to be extended across the existing Layer 3 network through the use of Layer 3 multipathing.

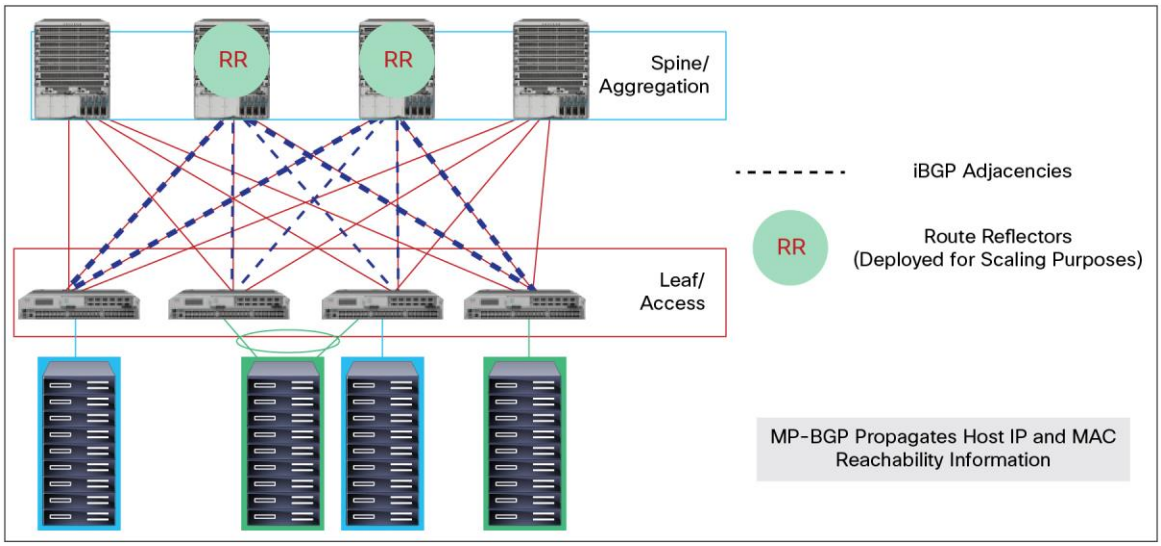
Figure 1. VXLAN Encapsulation



VXLAN IETF RFC 7348[1] specifies a multicast-based flood-and-learn mechanism, but customers who do not want to deploy multicast routing or who have scalability concerns related to flooding can use BGP as a control plane (Figure 2) to:

- Discover VXLAN tunnel endpoints dynamically
- Distribute attached host MAC and IP addresses and avoid the need for the flood-and-learn mechanism for unknown unicast traffic by prepopulating and disseminating the MAC addresses of the hosts
- Use a unicast network core (without multicast) and ingress replication for forwarding Layer 2 multicast and broadcast packets
- Terminate Address Resolution Protocol (ARP) requests early and avoid flooding

Figure 2. BGP-EVPN Control Plane for VXLAN



The BGP-based control plane for VXLAN is compliant with IETF drafts[2] that specify the BGP-EVPN control plane for overlays. It uses the Ethernet virtual private network (EVPN) address-family extension of Multiprotocol BGP to distribute the requisite overlay reachability information. While the Cisco Nexus 9000 Series is the first cisco platform, this control plane will be supported on other Cisco platforms[3].

Transparent Virtual Machine Mobility Support

The control plane supports transparent virtual machine mobility and quickly reconverges reachability information to avoid hair-pinning of east-west traffic. A top-of-rack (ToR) switch detects that a virtual machine has moved behind it by snooping on Domain Host Configuration Protocol (DHCP) or ARP packets. It populates the reachability information in BGP and advertises the updated MAC address route to its peers with an updated sequence number. When the original ToR switch receives the route update with the modified sequence number, it sends a withdraw message for the stale reachability information (Figures 3, 4, and 5).

[1] IETF RFC 7348 VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks
[2] draft-ietf-l2vpn-evpn, BGP MPLS based Ethernet VPN (EVPN)
draft-sd-l2vpn-evpn-overlay, A Network Virtualization Overlay Solution using EVPN
draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-03, IP Inter-Subnet Forwarding in EVPN
[3] On Cisco Nexus 9300 and 9500 platform switches, Cisco Nexus 7000 Series Switches (F3 Series only), and Cisco ASR 9000 Series Aggregation Services Routers

Figure 3. VXLAN Control Plane: Host Attaches

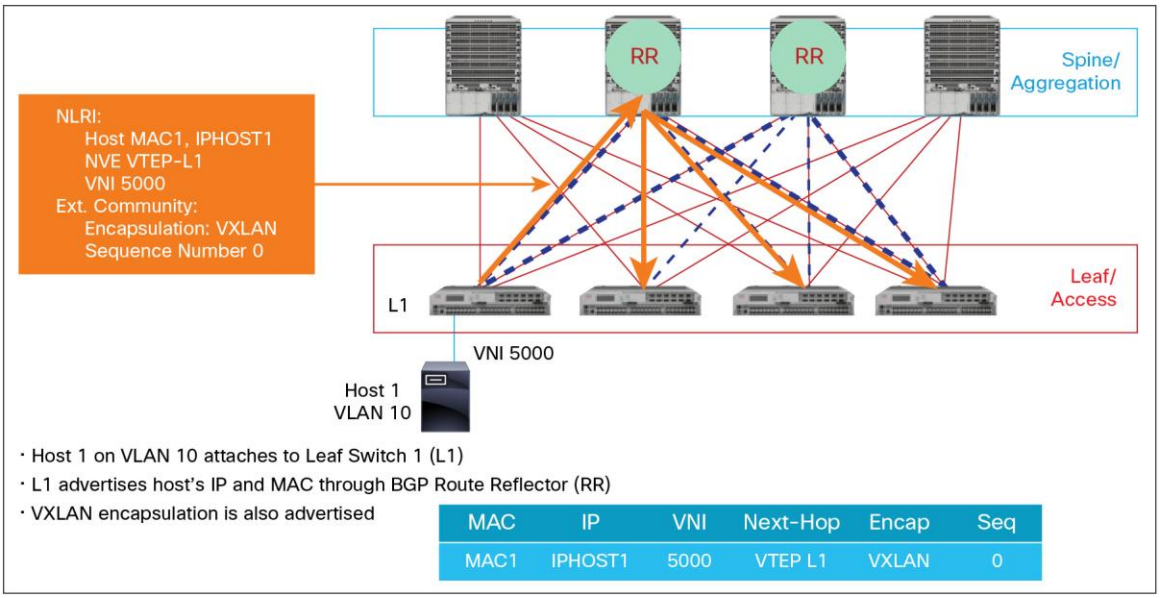


Figure 4. VXLAN Control Plane: Host Moves

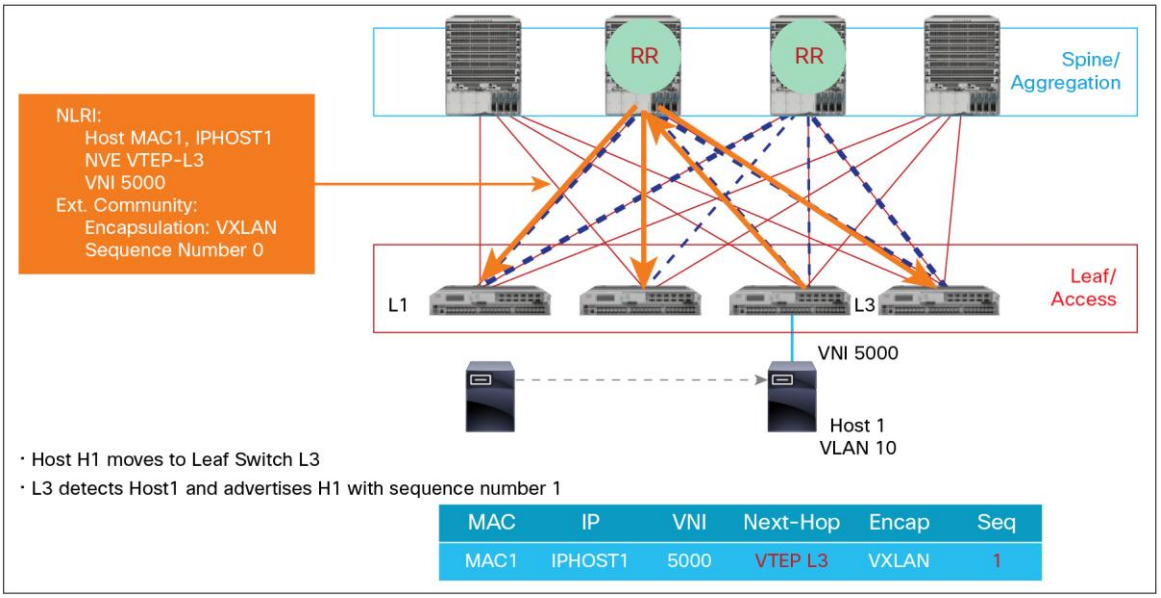
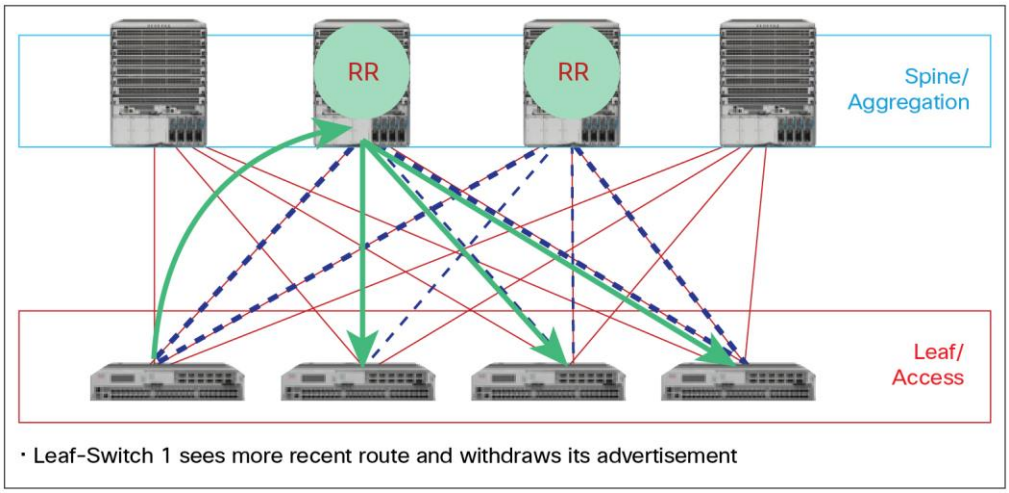


Figure 5. BGP Control Plane: Old Route Withdrawn



Layer 3 Forwarding at the Leaf Switches

In addition to traditional bridging behavior at the leaf switches, Cisco Nexus 9300 switches with the ALE ASIC offer the capability to route VXLAN overlay traffic at the leaf[4]. Unlike traditional Broadcom Trident II based platforms, which cannot VXLAN route the packet, the Cisco Nexus 9300 platform allows customers to bring their boundary between Layer 2 and Layer 3 overlays down to the leaf/access layer. Routing at the ToR layer facilitates a more scalable design, contains network failures, enables transparent mobility, and offers better abstract connectivity and policy.

Efficient Bandwidth Utilization and Resiliency with Active-Active Multipathing

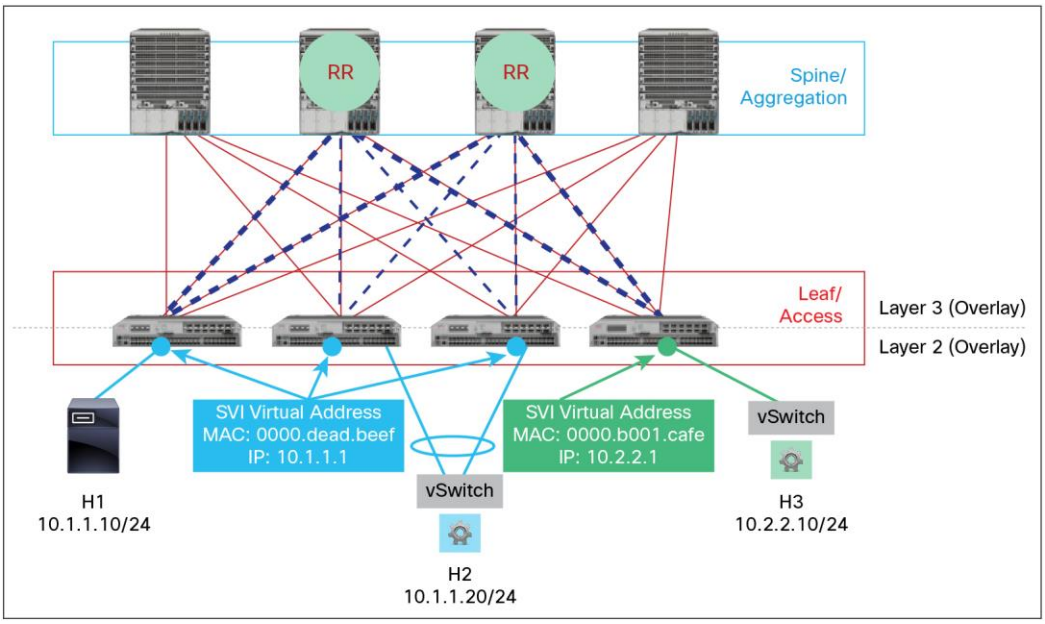
Cisco Nexus 9300 and 9500 platforms support VXLAN with virtual PortChannel in both multicast flood-and-learn and the BGP-EVPN control plane. This allows resiliency in connectivity for servers attached to access or leaf switches with efficient utilization of available bandwidth. VXLAN with virtual PortChannel is also supported for access to aggregation connectivity, promoting a highly available fabric.

Distributed Anycast Gateway

To facilitate optimal east-west routing while supporting transparent virtual machine mobility, leaf switches are assigned the same gateway IP and MAC address for each locally defined subnet. Having the same gateway IP and MAC address helps ensure a default gateway presence across all leaf switches. It removes the suboptimal routing inefficiencies associated with separate centralized gateways (Figure 6). It is enabled by the capability of the Cisco Nexus 9000 Series to VXLAN route the packets.

[4] Upcoming availability on Cisco Nexus 9500 platform switches, Cisco Nexus 7000 Series Switches (F3 Series only), and Cisco ASR 9000 Series

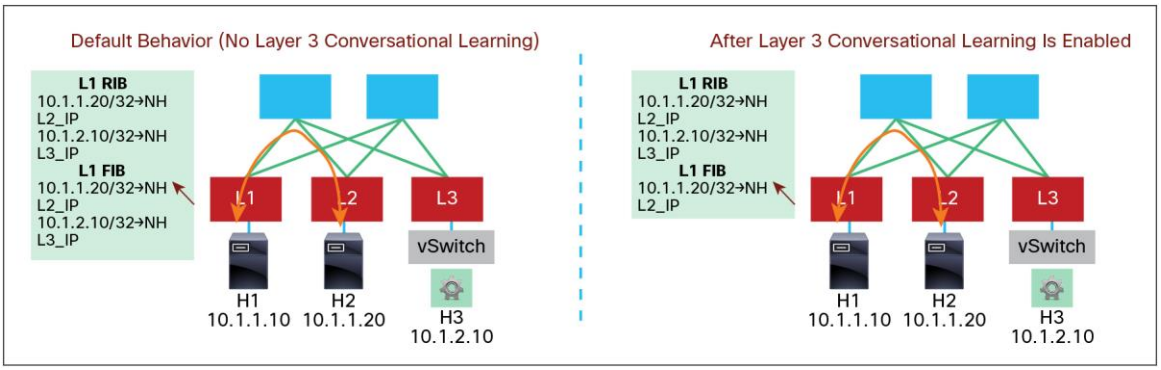
Figure 6. Distributed Anycast Gateway



Selective Download to Hardware Forwarding Tables for Conversational Learning

In a typical cloud deployment, leaf switches need to maintain a large number of MAC addresses and host prefixes in their hardware tables. As the data center grows, the problem is exacerbated, eventually forcing customers to replace their existing leaf switches and incur heavy capital investment. Conversational learning mitigates the hardware table size constraints on leaf switches and optimizes the number of host prefixes and MAC addresses that are stored in the hardware by storing only those host prefixes and MAC addresses for which a dual-sided conversation is detected. Although the routing information base stills learns all the host MAC address routes for a specific configured segment, the information is downloaded to hardware tables only when a conversation is detected (Figure 7).

Figure 7. Conversational Learning



Benefits of Cisco BGP EVPN Control Plane for VXLAN

- Removes the need for multicast flood-and-learn to enable VXLAN tunnels
- Supports VXLAN routing at the leaf switches to build a more scalable fabric
- Supports suppression of ARP and Neighbor Discovery Protocol (NDP) termination for unknown unicast destination addresses
- Supports Layer 2 active-active multipathing to promote efficient utilization of available links and greater resiliency
- Transparent support for physical and virtual hosts
- Transparent virtual machine mobility with optimized east-west routing
- Uses BGP, a proven and familiar protocol, that scales to support Internet-scale networks
- EVPN address family carries both Layer 2 and Layer 3 reachability information. This provides the flexibility to build either bridged overlays or routed overlays. While bridged overlays are simpler to deploy, routed overlays are easier to scale out
- Standards-based and interoperable with platforms that are consistent with the IETF draft
- BGP authentication and security constructs, which provide more secure multitenancy
- Rich BGP policy constructs, which provide policy-based export and import of reachability information. With these policy constructs, it is possible to constrain route updates where they are not needed and promote a more scalable fabric

Why Cisco?

The Cisco BGP Control Plane for VXLAN solution is a Cisco innovation that uses Cisco's proven track record for delivering scalable BGP solutions for multitenant data centers. The solution is standards based and interoperates with other solutions, giving our customers the flexibility to incrementally build a multitenant data center based on their needs. Cisco and its partners can help customers design and deploy a robust, dependable solution that addresses all aspects of deployment, operations, and optimization in a multitenant data center.

For More Information

Cisco Nexus 9000 Series Switches: Data Sheets and Literature

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.

Printed in USA C11-733737-00 01/15