Building scalable data centers with EVPN and VXLAN

Decoupling applications from the physical network enables enterprises and services providers to achieve application mobility, increased scalability, and better network
Juniper Technical White Paper

Draft 3

Contents

Using This Document

This document presents an overview of the EVPN/VXLAN architecture divided into two sections, each with its own purpose and audience.

Section I, EVPN/VXLAN Basic Concepts, is written for IT decision makers who are considering technology alternatives for next-generation data center designs. It presents key technical concepts in a business context to highlight the business benefits of EVPN/VXLAN. IT professionals including CIOs and executive vice presidents should find this information accessible and useful for planning the evolution of their data centers.

Section II, Technical Description of the EVPN/VXLAN Architecture, is written for network architects who are investigating alternatives for scalable data center networks. It goes into more technical detail to provide a basic technical understanding of important aspects of EVPN/VXLAN to inform additional research and vendor conversations.

SECTION I: EVPN/VXLAN BASIC CONCEPTS

1. Data Center Challenges Today

The current data center network is coming under pressure due to a number of major trends[1]:

- Cloud-based resources are becoming an increasingly larger part of the enterprise's IT strategy, requiring a network architecture that can accommodate cloud-based services without compromising security or performance.
- Agile software methods and DevOps are accelerating application development, putting pressure on the network to provide a high-bandwidth, low-latency platform.
- The demands of data center users for anytime, anywhere access and high levels of responsiveness are becoming harder and harder achieve with today's network architectures.

These trends are driving data center architects to re-envision the network with three key goals in mind:

- **Scalability:** Some enterprises are accommodating growth by increasing their use of cloud services, while others are deploying their own private and hybrid clouds. Service providers must grow rapidly to have sufficient capacity to meet demand. Legacy networks are often too rigid and difficult to change to support the scalability needs of the large enterprise and the service provider.
- **Operational flexibility**: As enterprises expand their geographic reach, they face problems relating to physical distance between data centers and users, as well as shrinking maintenance windows due to around-the-clock operations. The new data center network must support

[1] http://www.data centerknowledge.com/archives/2014/12/22/dynamic-data-center-3-trends-driving-change-2015/

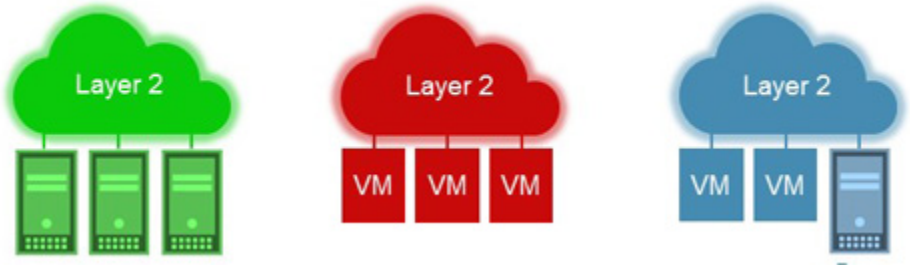
application mobility, allowing network administrators to easily migrate applications within the data center and between data centers for business continuity, maintenance without downtime, and load balancing.

- **High performance:** Today's users often complain about poor response and even outages of business-critical applications caused by bandwidth limitations and latency problems. Technologies such as multipathing and control plane learning can optimize network traffic flows, contain network faults, and ensure maximum utilization of bandwidth.
  VLANs: Not a complete solution

The primary problem with the legacy network is that applications are tied to the physical network topology, which has a number of implications:

- Application scalability is hampered by the network's inability to scale.
- Network managers cannot easily move applications within the data center or to other data centers.
- The rigid connection between applications and physical infrastructure makes it difficult to take advantage of cloud services.

In the traditional data center, network architects use virtual local-area networks (VLANs) to create Layer 2 logical networks that segregate users and applications to provide security in multitenant environments. VLANs also improve network performance by limiting broadcast traffic.



Layer 2 Logical Networks

However, this architecture is difficult to scale. The VLAN specification (IEEE 802.1ad) provides a relatively small address space which results in a maximum number of 4,096 VLANs .There is a one-to-one mapping between VLANs and logical networks, so the number of logical networks in the data center is also limited to 4,096. Multitenancy environments usually support a large number of users, each of whom may need multiple logical networks, so it's relatively easy to run up against this limit.

Another problem with the VLAN approach is that it constrains the movement of virtual machines (and thus the applications running on those VMs) to the physical hardware environment hosting the VLANs. Moving an application to another location in the data center or to another data center is a cumbersome and error-prone process; in practice, most network administrators avoid doing so unless absolutely necessary.

2. VXLAN Overview

This section provides an overview of the VXLAN protocol, including design goals, operational concepts, and limitations.

VXLAN Goals

The Virtual Extensible LAN (VXLAN) protocol is designed to address a number of limitations of the VLAN protocol[2].

Spanning Tree Protocol

Layer 2 networks often use the IEEE 802.1D Spanning Tree Protocol (STP) to avoid loops in the network. One of the drawbacks of this approach is that all traffic flows along a single path defined by the spanning tree, so alternate paths are blocked, even if they are more direct. The result is that the network never realizes its full capacity. In addition, the STP model does not support resiliency based on multipathing.

Cloud Environments

Cloud computing by its very nature is a multi-tenancy environment, meaning that the cloud service provider offers elastic, on-demand services to multiple customers over a shared physical infrastructure. Network traffic for each tenant must be securely isolated from the other tenants, which can be accomplished at either Layer 2 or Layer 3.

When the provider uses the Layer 2 approach, tenants are assigned one or more VLANs to provide the needed isolation and security. In these cases, the 4,096 limit on the number of VLANs severely hampers the provider's ability to utilize the shared infrastructure.

The Layer 3 approach also has significant problems. One tenant can use Layer 3 addresses within its networks which overlap the addresses of other tenants, requiring the cloud provider to provide another form of isolation. In addition, requiring the use of IP prevents individual tenants from communicating between VMs using direct Layer 2 or non-IP Layer 3 protocols.

Expansion across PODs

Service provider data centers are often organized by PODs, each of which consists of one or more racks of servers plus the associated networking and storage resources. If a tenant needs to expand beyond the resources of a single POD, the service provider needs to provision unused VMs from other POD. To do so requires stretching the Layer 2 environment across physical servers, which is difficult to do when using VLANs.

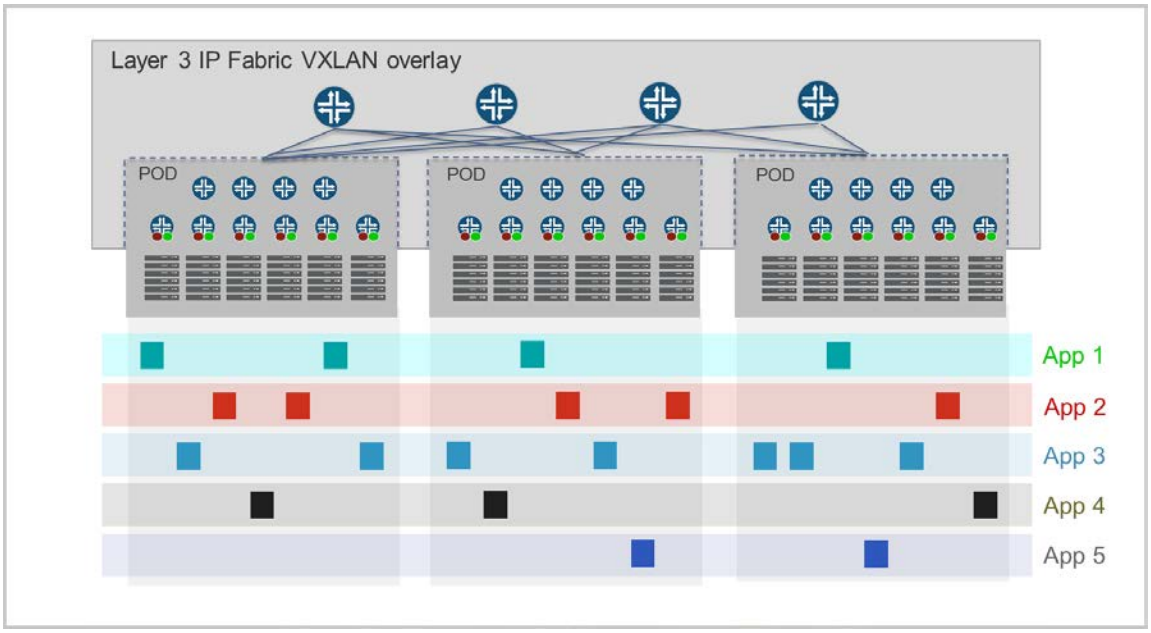
VXLAN Operational Concepts

The Virtual Extensible Local Area Network (VXLAN, standard IETF RFC7348) takes a major step in resolving the limitations of VLANs described above. VXLAN supports transport of native Ethernet

[2] This section draws heavily on the VXLAN spec RFC 7348, Mahalingam, et al., August 2014

packets inside a tunnel encapsulation[3]. VXLAN overlays a logical layer 2 network on a physical layer 3 network to allow the layer 2 network to span physical servers and enable scaling of the underlay network without impacting the logical networks in the VXLAN overlay.

VXLAN has a 24-bit addressing space for virtual networks, which allows for 16 million logical networks, effectively eliminating the VLAN addressing restriction. Implemented in hardware, VXLAN has become the de facto standard for overlays terminated on physical switches and is hardware-supported in many of the top networking chipsets, including Broadcom Trident 2 (used in Juniper QFX5100 switches), Juniper Q5 (QFX10000 switches), and Juniper Trio(MX routers)



VXLAN showing application mobility

Since VXLAN supports the transport of Layer 2 frames over a Layer 3 infrastructure, tenants are no longer constrained by the boundaries of any particular segment and can exist anywhere in a POD or data center and even across data centers. One of the biggest advantages of VXLAN is to untie tenants and applications from the physical networks and have the logical network (with VXLAN) be present anywhere in the network. In essence, by leveraging VXLAN, we no longer think of a network scale or a segment scale in the context of a single POD or a single data center.

VXLAN Limitations

VXLAN resolves a number of problems associated with VLANs but has limitations of its own, especially when it comes to the control plane. In fact, the first VXLAN draft relies on endpoint reachability using a

[3] A variant of VXLAN known of VXLAN-GPE supports the transport of non-native Ethernet packets within the VXLAN encapsulation. VXLAN-GPE is particularly of interest to data center architects for transporting routed IPv4/IPv6/MPLS packets. While it has uses in individual applications, VXLAN-GPE is not a recognized standard and is not widely supported by top networking vendors.

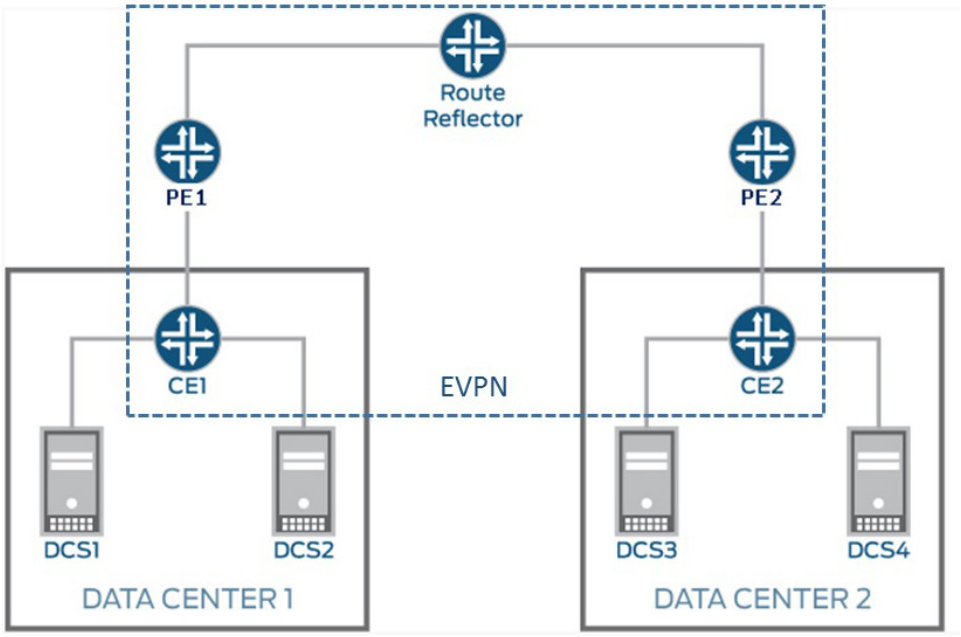
flood-and-learn mechanism by enabling multicast in the physical network or an underlay and associating a VXLAN ID or VNID with a multicast group

In short, VXLAN is a workable solution for the data plane but cannot offer the scalability, flexibility, and resiliency that service providers need in the control plane. That's where the Ethernet VPN (EVPN) comes in.

3. MP-BGP and EVPN: Control Plane Solution for VXLAN networks

Just over a year old, the EVPN draft standard is steadily becoming the control plane solution of choice for data center VXLAN networks. An EVPN is essentially a Layer 2 virtual bridge that can be used to connect dispersed sites.

As with other types of VPNs, an EVPN is comprised of customer edge (CE) devices (host, router, or switch) connected to provider edge (PE) devices (see figure).



Two data centers connected by an EVPN

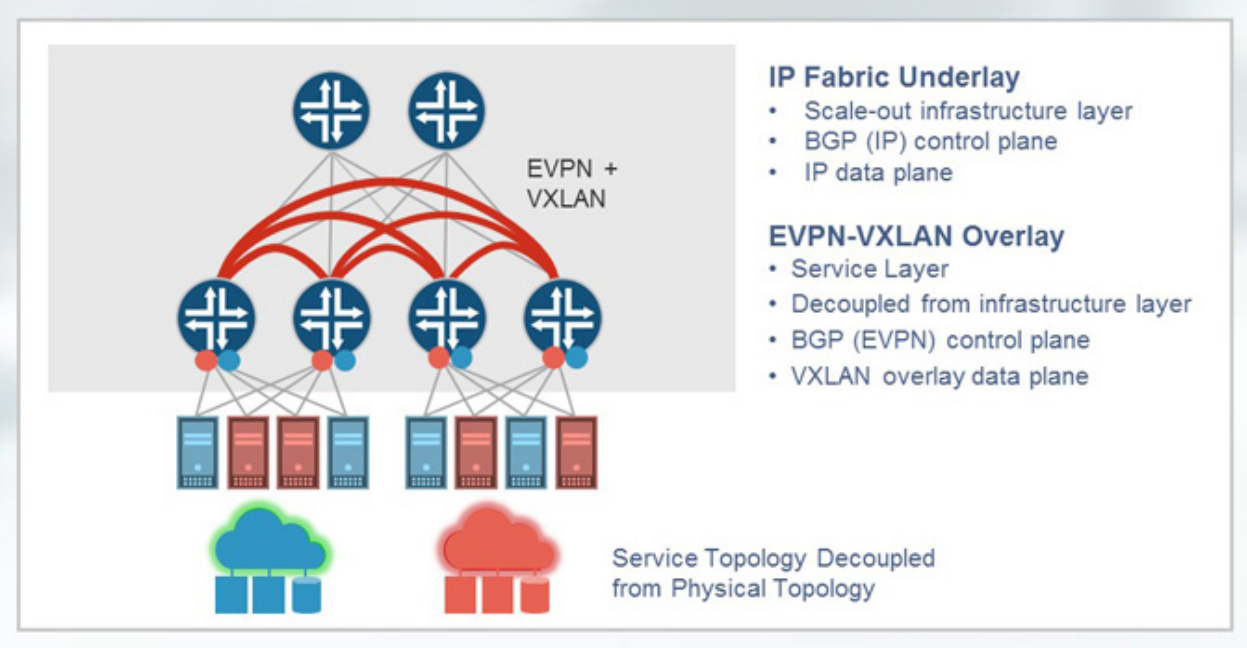
MPLS or VXLAN for Transport?

In its first deployments, EVPN often used the MPLS protocol for transport. More recently, EVPN has evolved to use VXLAN as the transport protocol in addition to MPLS[4]. VXLAN encapsulation and forwarding functions are essentially identical to those of MPLS, plus VXLAN has features designed specifically for the workloads within the data center.

[4] For a discussion of the benefits of VXLAN over MPLS, see "Overlay Networking & VXLAN Means MPLS in the Data Centre is Dead," by Greg Ferro, December 17, 2013. http://etherealmind.com/overlay-networking-vxlan-means-mpls-in-the-data-centre-is-dead/

BGP for Control Plane

The EVPN protocol uses the control plane infrastructure of the Border Gateway Protocol (BGP), a routing protocol with years of development that is widely regarded as the de factor protocol for the Internet[5]. An extension known as Multiprotocol BGP (MP-BGP) provides multiprotocol support lacking in the original BGP standard. MP-BGP enables BGP to carry both Layer-2 MAC and Layer-3 IP information at the same time[6].



EVPN/VXLAN Overlay with IP Fabric

4. Benefits of the EVPN/VXLAN solution

Combining VXLAN transport with the BGP/EVPN control plane creates a solution that has significant advantages over existing approaches when it comes to designing the next generation of scalable, multitenancy data centers.

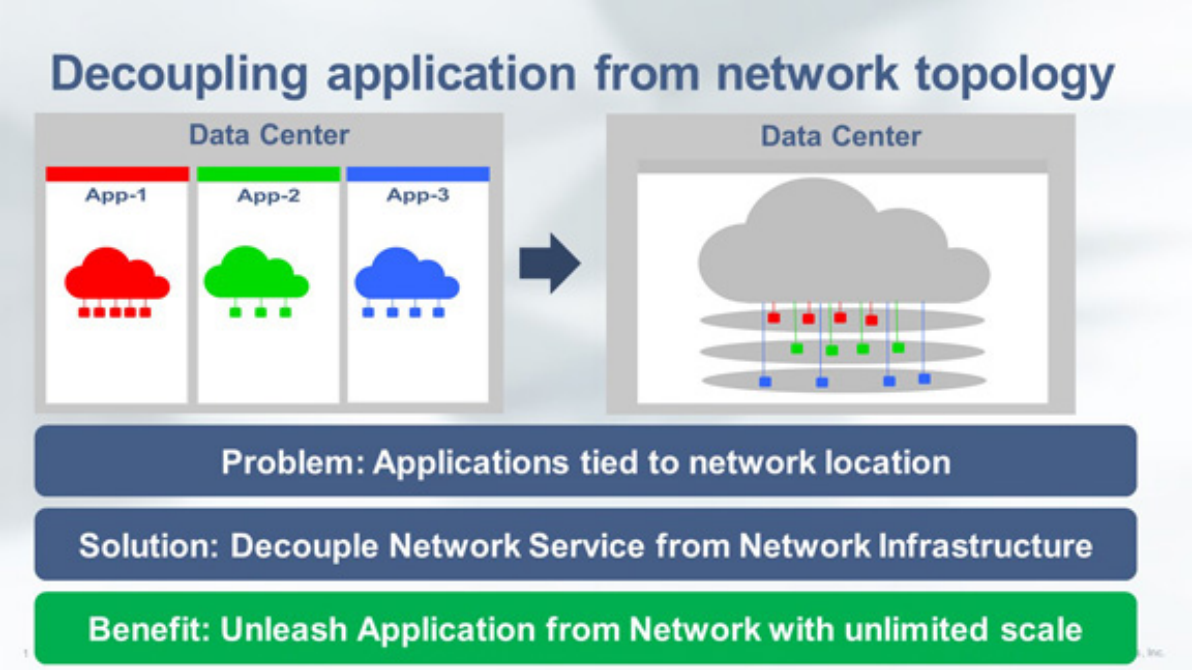
[5] A good overview of BGP can be found in the article, "BGP essentials: The protocol that makes the Internet work," by Ivan Pepelnjak, November 2007. http://searchtelecom.techtarget.com/feature/BGP-essentials-The-protocol-that-makes-the-Internet-work.

[6] The term "BGP" in the rest of this paper is assumed to include the multiprotocol extension.

Application Mobility

The EVPN/VXLAN solution provides complete end-to-end mobility for applications hosted on virtual machines, which enables a range of operational benefits such as:

- Application uptime. When network maintenance is required, the network manager can live-migrate applications to other servers in the data center or even to another data center without impacting availability.
- Business continuity. In the case of an anticipated disaster event, business-critical applications can be proactively migrated offsite until the danger is passed.
- Workload balancing. The ability to easily migrate applications between data centers allows network managers to balance traffic across multiple sites. Moving applications closer to users improves performance and minimizes WAN traffic.
- Facilities flexibility. Activities such as data center consolidation, large-scale facility expansion, or wholescale migration between facilities used to be extremely disruptive to users. Now network architects can minimize these effects with phased migrations that virtually eliminate application downtime.



EVPN/VXLAN architecture decouples applications from the network topology

Scalability

As networks grow larger, relying on the data plan for learning host reachability is neither scalable nor reliable. Using a scalable protocol like BGP EVPN in the control plane for end point discovery offers a scalable solution for use cases that within and across data centers. BGP has a long history in large IP networks, and with the EVPN extensions can also support Layer 2 networks by advertising mac reachability information.

Resiliency

As more and more business-critical applications are hosted on cloud services, it becomes increasingly important for the network to recover quickly from a network path failure. EVPN includes the mass MAC withdrawal feature that allows devices on an EVPN/VXLAN network to immediately signal and switch to a new path in the event of a failure.

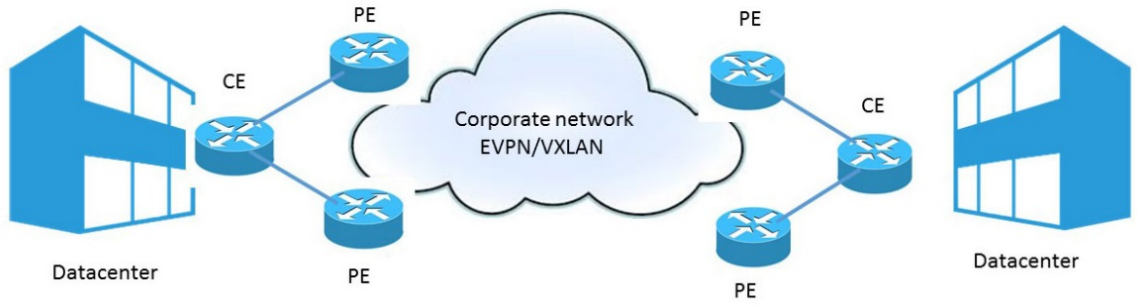
Service Continuity

The multihoming feature of EVPN provides a redundancy mechanism that allows two or more PE routers to be connected to the same CE device. In the event of a PE router or PE-CE link failure, another PE router can provide network service to the customer site without manual intervention.

Rich Policy Control

EVPN allows more fine-grained control through the use of export and import policies, functionality that is similar to Layer-3 VPN route policies. Users can deploy this feature with an SDN controller and use VNID for route filtering and segmentation of virtual networks.

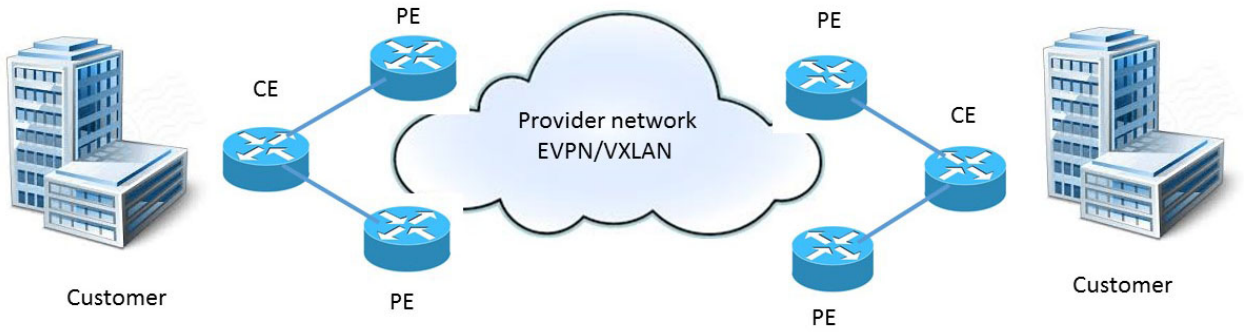
5. Use Case: Datacenter Interconnect



Datacenter interconnection presents a host of challenges such as workload mobility and business continuity that past solutions have only partially addressed. Maximizing application availability and network resource utilization requires load balancing, all-active redundancy, and fast convergence. Scalability of MAC addresses is also paramount due to the proliferation of virtual machines.

The EVPN/VXLAN solution enables scalable Layer 2/Layer 3 interconnections between virtualized data centers. Applications can move freely between data centers without disrupting forwarding databases (FDBs). The EVPN multihoming capabilities allow network architects to design the network infrastructure for high application availability. EVPN/VXLAN also improves network performance by eliminating multicast flooding on the data plane.

6. Use Case: Service Provider Infrastructure



Service providers can take advantage of the EVPN/VXLAN network architecture to offer integrated Layer 2 and Layer 3 services over a single interface per customer. This solution supports all-active forwarding and load balancing to improve network efficiency. The use of MP-BGP as the control plane protocol offers scalability and reduces or even eliminates flooding. EVPN/VXLAN removes the need for multiple VPN protocols, which simplifies network management and provisioning. MAC/IP provisioning can be accomplished programmatically from a network management system database.

SECTION II: TECHNICAL DESCRIPTION OF EVPN/VXLAN ARCHITECTURE

7. EVPN Overview

While there are many protocols that advertise and exchange Layer 2 MAC addresses, OVSDB (Open Virtual Switch Database Protocol) and EVPN (Ethernet VPN) are the most popular ones deployed in the data center today. The OVSDB protocol currently is only used by VMware NSX controllers. EVPN is used in controllerless environments as well as Juniper Contrail and Nuage controllers. EVPN has traditionally used MPLS for transport, however, it is being replaced by VXLAN in many new deployments. EVPN is standards based and supported on a number of software-defined network (SDN) controllers, which makes it a good choice for organizations who are planning their move to SDN.

The MAC Addressing Challenge

Layer 2 switches use the flood-and-learn mechanism to maintain up-to-date MAC addresses in their routing tables. This approach is well-suited to local-area networks but has significant problems for Layer 2 data center fabrics and data center interconnections. The most serious is uncontrolled flooding, which can consume a significant percentage of network capacity and cause slowdowns or even outages. Uncontrolled flooding can be prevented with the spanning tree protocol (STP) and other protocols, but these approaches add complexity to network operations and reduce utilization levels of network components. In the case of interconnected data centers, a network problem in one data center can propagate to the others.

Flood-and-learn also has scalability problems for both physical and virtual networks. Physical scaling is difficult due to spanning tree loop issues that cause certain links to be disabled, leading to network

inefficiency and resource underutilization. Additionally, the highly static and time-consuming deployment causes significant overhead in terms of ascertaining the number of links, capacity increases, and other factors. Logical scaling is limited by the VLAN restriction of 4,096 VLAN addresses. Workarounds such as Q-in-Q or 802.1ad can overcome the addressing issue but are laborious to provision and maintain. Hosting providers who use bare-metal servers generally employ flood-and-learn and avoid spanning tree protocols altogether with MC-LAG or Virtual Chassis.

How EVPN Addresses VXLAN Challenges

EVPN solves these problems by introducing MP-BGP as the control plane. In EVPN, MAC addresses are treated as routes in the BGP table. The table entries can be a MAC address only or a MAC address plus an IP address, which is needed for Address Resolution Protocol (ARP). VLAN tags can be included as well.

Scalability

Because EVPN is based on BGP, which currently supports millions of Internet prefixes, EVPN can support millions of host and MAC addresses using a new RLI called EVPN network layer reachability information (NLRI).

Virtual Machine Mobility

EVPN is designed for fast convergence, which facilitates the movement of virtual machines within a data center or even across data centers. It also includes a mechanism to avoid duplicate MAC addresses caused by moves, eliminating the so-called MAC flapping problem.

Network Utilization

EVPN makes forwarding decisions based on the MAC addresses that are stored as BGP routing table entries. Unlike protocols that define a single path between routers, EVPN supports multiple active paths for all traffic except for broadcast, unknown unicast and multicast (BUM). This approach avoids the problem of unused links and maximizes resource utilization.

Fast Failure Detection

Given that cloud services are hosting more and more business-critical applications, redundancy for services is provided at the application level. Therefore, it is paramount to detect failures and flush stale state information. EVPN provides the mass MAC withdrawal feature that allows devices on an EVPN network to immediately signal and switch to a new path.

Rich Policy Control

EVPN allows finer control over the advertised prefixes by making use of export and import policies. EVPN provides similar functionality to Layer 3 VPN route policies. Network operators can deploy this feature with an SDN controller and use VNID for route filtering and segmentation of virtual networks.

8. EVPN Concepts

EVPN introduces new terminology, a series of new route types, and new extended communities.

Terminology

Three new terms are important for understanding the descriptions that follow:

Designated Forwarder (DF)

A DF is chosen based on the Ethernet segment route. DF election is necessary when a CE is multihomed to two or more PEs. In single-active mode, the non-DF does not send traffic to the access interface that is connected to the multihomed CE. Also, the DF is the device that generates the split horizon label.

Ethernet Segment ID (ESI).

Similar to DF, the ESI concept is used in EVPN to support multi-homing. An Ethernet segment consists of one or many interfaces on different PEs through which a particular multi-homed CE can be reached. From a multi-homed CE perspective, the different interfaces in an Ethernet segment appear as regular child links in a LAG interface. The ESI value should be unique for all the EVPN instances across all the EVPN PEs.

Ethernet Tag Identifier (ETI).

In an EVPN/VXLAN network, the ETI carries VNID information. VNID is used in a similar way as VLAN-ID for customer segmentation, but scales better than VLAN because it is a 24-bit long designator.

Route Types

Given the new EVPN NLRI and the additional challenges of supporting active/active in multihoming scenarios, EVPN introduces four route types.

Route Type-1: Ethernet Auto-Discovery Route

Advertised on per-EVI (EVPN Instance) and per-ESI basis (Ethernet Segment Identifier). Ethernet auto discovery routes are required for deployments with multihomed CEs. EVPN allows auto-discovery per Ethernet segment or EVPN instance.

Per-Ethernet auto-discovery indicates all-active or single-active multi-homing capabilities. Single active is equivalent to active/standby mode, while all-active is equivalent to active/active mode.

Per-EVPN Instance auto-discovery, also known as aliasing, allows the ingress PE to load balance towards multiple egress PEs that are connected to the same multi-homed CE.

Route Type-2: MAC/IP Advertisement Route.

Allows the endpoint IP and MAC addresses to be advertised within the EVPN NLRI, enabling control plane learning of endpoint MAC addresses. JUNOS uses the same label for MAC/IP advertisement route as in Ethernet AD per EVI route leading for optimization in label usage. EVPN can advertise labels per MAC, ESI or EVI, offering flexibility in determining de-multiplexing parameters.

Route Type-3: Inclusive Multicast Route.

Sets up a path for BUM traffic to the remote PE by VLAN or ESI. It allows PE to send BUM traffic from a CE on a VLAN in an EVI to all the other PEs that span that VLAN in that EVPN instance. The point to multi-point path to create a tree from the source PE of the BUM traffic to the receivers PE can be created using existing and well-defined constructs such as P2MP tunnels, which are optimized for sending multicast traffic through the core. EVPN also supports ingress replication in which the ingress PE itself replicates and sends copies of the BUM traffic to the receiver PE.

Route Type-4: Ethernet Segment route.

Allows the CE to be multihomed to two or more PEs—in single/active or active/active mode. PEs connected to the same Ethernet segment discover each other through the Ethernet segment route. The Ethernet segment route is used in conjunction with the ES-Import extended community (see below) which ensures that only PEs with the same ESI can import this route. DF election is carried out based on ES route and on a per-EVI basis which helps in load sharing of traffic across multiple PEs.

Extended Communities

EVPN also provides for new extended communities.

ESI Label Extended Community

A new transitive extended community advertised with the Ethernet AD route. This label enables the split-horizon mechanism for multihomed CEs.

MAC Mobility Extended Community

A new transitive extended community advertised with the MAC/IP route. MAC mobility or MAC move is defined as a MAC moving from one Ethernet Segment to another. These Ethernet segments may reside on the same PE or different PEs.

Default Gateway Extended Community

Associating this extended community with a MAC/IP route indicates that route as a default gateway. This is the default behavior when an IRB interface is configured in JUNOS. Advertising community helps reduce traffic flooding across the network and optimize egress VM traffic.

9. EVPN Enhancements for VXLAN

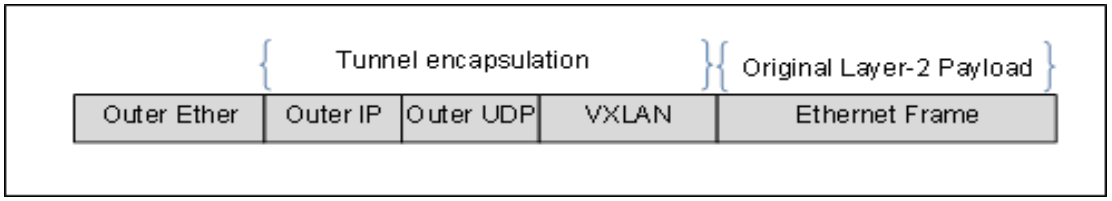
As noted earlier, the trend in EVPN deployments is to use VXLAN instead of MPLS for the transport function. The biggest difference between the two is that VNIDs are exchanged instead of MPLS service labels. Enhancements within the EVPN protocol standardize support across different platforms and vendors.

BGP Encapsulation

When EVPN is used as the protocol for Network Virtualization Overlay (NVO), EVPN advertises its encapsulation capabilities through the BGP encapsulation extended community[7]. To support the EVPN encapsulation, all EVPN PEs define the tunnel type as VXLAN for next hop encapsulation through the BGP community. The egress PE attaches the BGP encapsulation extended community attribute with tunnel encapsulation type VXLAN to EVPN MAC routes, the EVPN inclusive multicast route, and EVPN AD per-EVI routes. If ingress PE and egress PE do not have a common encapsulation type, the egress PE ignores BGP EVPN routes advertised by the ingress PE.

Packet Format and Tunnel Creation

In EVPN/VXLAN solutions, the Layer 2 payload is directly encapsulated with the VXLAN header at the ingress PE. At the egress PE, the VXLAN header is popped to further process the original Layer 2 payload.



VXLAN Header at Egress PE

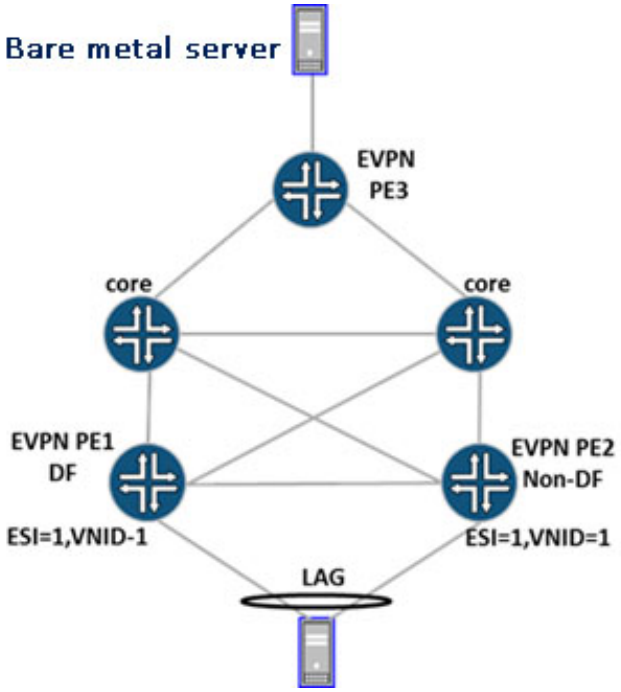
In EVPN/VXLAN solutions, the EVPN control plane is responsible for the discovery of the remote VTEPs (VXLAN tunnel endpoints). Once the two tunnel endpoints are discovered, a dynamic VXLAN tunnel is created automatically between the logical source and remote VTEP interfaces.

Locality Bias

Locality bias is used to implement split horizon. Because there are no service levels as in MPLS, the split-horizon forwarding rule is modified to use the IP address of the EVPN PE instead of the MPLS ESI label. Each EVPN PE tracks the IP address of the remote multi-homed PE as the source VTEP address for each VXLAN packet received from the PE. Remote PEs are configured with the same ESI for one or more EVIs.

The local bias forwarding filtering rule is enforced on both ingress and egress PEs for the multi-destination traffic. The figure below shows bare metal servers (BMS) connected to EVPN PE devices, which can be switches or routers capable of supporting VXLAN encapsulation and EVPN with VXLAN in the control plane. EVPN has locality bias support for egress and ingress PEs.

[7] In JUNOS, by default a BGP protocol *next-hop* is resolved through label based protocols like RSVP, LDP, and BGP-LU. However, in EVPN with VXLAN overlay, this mechanism is overridden.



EVPN Traffic Pattern at Egress PE and Ingress PE

Aliasing

In the figure above, the bare metal server is multihomed to the EVPN PE. BUM traffic is load-balanced and sent over one of the links. Based on the hash result, BUM traffic might go to PE2 (non DF), in which case the server's MAC address is learned by the rest of the PEs with PE2 as the next hop. To support load-balancing or aliasing from PE3, PE1 advertises per-EVI AD route, which establishes PE1 as a valid next hop at PE3. This procedure for EVPN with VXLAN transport is similar to the normal EVPN with MPLS as transport.

10. Layer 3 Routing across Layer 2 Logical Networks

While VXLAN extends the Layer 2 domain, the problem of doing Layer 3 routing across these extended VXLAN segments requires a different solution to three Layer 3 routing problems: host routing, default gateway routing, and prefix routing.

Host Routing

The EVPN protocol supports dissemination of MAC routes using EVPN Type-2 routes. The same route can optionally also carry the host IP route.

Proxy ARP

When the host IP route is carried with the host MAC route and the BGP route target equals the VXLAN segment, all other PEs on the same VXLAN segment receive the host IP route. Using this route, the PE can create a database of host-IP to host-MAC mapping. When another host in the same segment sends and ARP request for the host-IP route, the PE closest to the requesting host can respond with the ARP response, which avoids flooding of the ARP request.

Shortcut Routing

When a host IP route is learned via EVPN, the other PEs in the host segment can directly route the packet to the host PE without having to go to the default router. In this case, all other PEs will have a more specific host IP route in addition to the subnet route that points to the default router of the subnet. The ingress PE also performs the Ethernet MAC rewrite for the host using the VNI and host-MAC present in the EVPN type-2 route.

Shortcut routing is useful when the overlay routing is done multiple hops away from top of rack switches or if the PE in the overlay segment routing layer cannot support IRBs for all the VXLAN segments in the POD. In either case, if the packet were to be routed via the default gateway, it will require extra hops for it to go through the network.

Reducing Next-hops with Routed-VNI

While shortcut routing has a positive impact on bandwidth usage, it also has a serious drawback: Every host route and its next hop rewrite is present on all PEs. Host next hops are a scarce resource in most ASICs, so this can be a serious issue.

In the VXLAN overlay, the egress PE, not the ingress PE, does the host Ethernet header rewrite. The egress PE must identify the VRF in which to perform the lookup. The ingress PE can signal this information to the egress PE by use of a routed VNI which is unique to each VRF. The ingress PE performs the Ethernet header rewrite for all hosts in the VRF. Using routed VNI reduces the number of next hops as much as three orders of magnitude.

Default Gateway Routing

In an EVPN/VXLAN environment, the host default gateway may have to be present on multiple PEs. There are two mechanisms to ensure that all PEs route packets that are destined for the routed-MAC of any of the PEs.

Anycast Virtual MAC

In this mechanism, all PEs must be configured with the same Anycast virtual MAC address. As a result, all PE can route packets destined for a router MAC. This mechanism works well within PODs or data centers as long as the PEs can be properly configured.

Default-gateway MAC

By using an EVPN Type 2 Route, a gateway MAC is made known to all other gateways in the same Layer 2 segment. This is the preferred mechanism when the Anycast Virtual MAC mechanism cannot be implemented.

Prefix Routing

Prefix routing across PEs in a data center is accomplished using a routed VNI mechanism similar to the vrf-table-label option, the only difference that VNIs are allocated per VRF. If the routed VNI is same across all PEs for a given VRF, the BGP route target for the VRF can be derived using the routed-VNI. If the routed VNI is not the same on all PEs, route-targets must be configured individually.

11. Layer 2 Datacenter Interconnect

Technologies such as VPLS over MPLS were sometimes used in the past to extend Layer 2 across sites. With the growing trends of multi-tenancy and data center automation, this kind of solution is becoming more compelling. The inter-DCI mechanisms in EVPN environments use many of the same concepts as VPLS over MPLS.

To interconnect NVEs in different PODs and allow the NVEs to use VNIs as globally unique identifiers within the data center, a gateway must be employed at the edge of the data center network to translate VNIs across network boundaries. In this case, the VNIs are globally unique within a POD but not necessarily across the PODs. This section describes options for data center interconnection (DCI), all of which assume that VNIs are globally unique within a data center and that each data center has a gateway. One of the goals of this DCI method is to ensure that NVEs inside the data center only need be aware of a single VNI space and the gateways handle the complexity of managing multiple VNI spaces.

To provide a scalable solution, gateways should not have to perform route lookup for DCI traffic. If the gateways cannot support this requirement, then an acceptable alternative is for the gateway to maintain the routes in its data center but not have to maintain routes from remote data centers in its forwarding information base.

VNI Translation

When interconnecting data centers that use different VNIs to represent the same VLAN, VNI/VSID translation functionality is needed at the data center gateways.

Assume two interconnected two data centers, DC1 and DC2. Gateways at the edge of each data center translate VNIs/VSIDs into an intermediate value, sometimes called the DMZ VNI. There are two options for the DMZ VNI, both of which must be agreed upon by the operators of DC1 and DC2:

- Assign a single global DMZ VNI to each VLAN: The route-target can then be auto-derived from the DMZ VNI.
- Assign the DMZ VNI downstream: In this case, the operators must agree on the route-targets to be used at the gateways.

Crossing between autonomous data centers always requires VNI translation and may require IP address translation if NVE addresses are private. As a result, gateways must be aware of multiple domains in the VNI space.

BGP Updates

When BGP updates are sent to the remote gateway, there are two options for VNI allocation and route target handling:

- Translate VNIs using the DMZ VNI described earlier: The route targets can be auto-derived from the DMZ VNI.
- Do not translate VNIs: In this case, the Ethernet tag in the update message does not change. The route targets and the import/exports need to be configured correctly on both the gateways to process the remote BGP routes.

12. Multicasting in Overlay Networks

Multicasting in overlay networks presents several challenges. The multicast control plane not only has to learn the overlay end-points, but must correlate this information to the underlay replication mechanism which transports information to the end-points. The first section discusses the various mechanisms available to make copies in the underlay network, while the later sections describe how Layer 2 and Layer 3 multicast is implemented in the control plane to use these mechanisms.

Replication

The underlay network can send copies of the data to relevant end-points or it can be made completely agnostic of overlay multicast. The latter keeps the underlay free of multicast protocols and scales much better in the control plane.

Underlay Multicast

To send copies of the data to each egress PE, underlay routers use the underlay multicast group. The underlay multicast group can be assigned by overlay segment (VNI) or overlay multicast group.

- Overlay segment: The underlay multicast group is configured in the underlay Layer 3 instance per overlay customer bridge-domain. The underlay multicast group is used to deliver all BUM data traffic in the segment. One disadvantage of this scheme is extra flooding for all overlay multicast groups. This is the scheme proposed in the base VXLAN draft.
- Overlay multicast group: One or more overlay-multicast-groups are configured to map to each underlay multicast group. Using this information, egress PEs can join the underlay multicast group corresponding to the overlay-multicast-group for which they receive a multicast join.

Multicast-free Underlay

Due to the complexity of multicast protocols, many providers avoid running multicast in the underlay altogether. In this case, the underlay must carry unicast copies of the overlay multicast data, one copy for each egress PE that is interested in the multicast overlay group. There are two primary methods for replicating the data, ingress replication and assisted replication.

- Ingress replication: The ingress PE sends a unicast copy of data to each egress PE that needs the data.
- Assisted replication: A replication PE is designated to send data on behalf of the ingress PE. The PE hosting overlay gateway is a good candidate for this approach because the gateway receives overlay IGMP reports about overlay routers in the segment and use this information to create the replication list. The ingress PE learns about the assisted replication either through configuration or EVPN signaling.

Layer 2 Multicast

IGMP snooping allows a network switch to listen to communications between hosts and routers and use this information to determine the destination links for each multicast stream. IGMP snooping creates a replication tree for both intermediate hops and end nodes. In overlay networks, the intermediate hops

(P routers) simply pass along VLAN traffic and thus have no knowledge of VLAN groups. IGMP snooping can be implemented using two mechanisms.

Overlay Edge

In this mechanism, IGMP snooping occurs at the overlay edge. In the underlay, multicast data is flooded to all PEs in the overlay segment VNI using the underlay multicast group. The edge PE only floods overlay data to the relevant ports at the edge using the IGMP snooping information for the overlay multicast group.

EVPN Type-2 Route

Using an out-of-band control plane helps optimize multicasting in the core. With EVPN, the overlay multicast-group membership snooped via IGMP at the edge can be signaled using EVPN type-2 routes for multicast groups. This information is propagated to all PEs in the overlay segment using EVPN route-target for the overlay segment. There is no need for a data-plane flooding mechanism. To review core routers of multicast support, the edge PE can maintain the replication list and send unicast copies using either of the replication mechanisms described earlier.

Layer 3 Multicast

Layer 3 multicast in an environment with multiple overlay segments per user poses some of the same challenges as Layer 3 unicast support. This section presents several alternative approaches.

Unoptimized Layer 3 Multicast

In this approach, the traditional Layer 3 multicast is performed by the gateway routers. The overlay network only supports Layer 2 multicast as described in the previous section. While this approach is simple to deploy, the disadvantages are 1) multiple copies of data are sent through the network core, 2) gateways must run a traditional multicast protocol such as PIM on the overlay segments in order to perform designated router election, and 3) traffic can trombone at the sender and receiver DR gateways.

Optimized Shortcut Layer 3 Multicast with Routed-VNI

Layer 3 multicast support overcomes the shortcomings of unoptimized Layer 3 multicast. In this approach, EVPN type-2 route signals the overlay multicast group interest using routed-VNI. The ingress PE then sends a unicast copy to the egress PE.

13. Summary/More Information

VXLAN is an overlay technology that encapsulates MAC frames into a UDP header at Layer 2. Communication is established between two virtual tunnel endpoints (VTEPs). VTEPs encapsulate the virtual machine traffic into a VXLAN header, as well as strip off the encapsulation. Virtual machines can only communicate with each other when they belong to the same VXLAN segment. A 24-bit virtual network identifier (VNID) uniquely identifies the VXLAN segment. This enables having the same MAC frames across multiple VXLAN segments without traffic crossover. Multicast in VXLAN is implemented as Layer 3 multicast, in which endpoints subscribe to groups.

EVPN is a flexible solution that uses Layer 2 overlays to interconnect multiple edges (virtual machines) within a data center. Traditionally, the data center is built as a flat Layer 2 network with issues such as flooding, limitations in redundancy and provisioning, and high volumes of MAC addresses learned, which cause churn at node failures. EVPN is designed to address these issues without disturbing flat MAC connectivity.

In EVPN, MAC address learning is driven by the control plane, rather than by the data plane, which helps control learned MAC addresses across virtual forwarders, thus avoiding flooding. The forwarders advertise locally learned MAC addresses to the controllers. The controllers use MP-BGP to communicate with peers. The peering of controllers using BGP for EVPN results in better and faster convergence.

With EVPN, MAC learning is confined to the virtual networks to which the virtual machine belongs, thus isolating traffic between multiple virtual networks. In this manner, virtual networks can share the same MAC addresses without any traffic crossover.

To learn more about Juniper solutions for the data center, please visit www.juniper.net/data center