INTERNET-DRAFT
Intended Status: Informational

Sami Boutros
VMware

Ali Sajassi
Samer Salam
Dennis Cai
Samir Thoria
Cisco Systems

Tapraj Singh
John Drake
Juniper Networks

Jeff Tantsura
Ericsson

Expires: September 17, 2016 March 16, 2016

VXLAN DCI Using EVPN
draft-boutros-bess-vxlan-evpn-01.txt

Abstract

This document describes how Ethernet VPN (E-VPN) technology can be used to interconnect VXLAN or NVGRE networks over an MPLS/IP network. This is to provide intra-subnet connectivity at Layer 2 and control-plane separation among the interconnected VXLAN or NVGRE networks. The scope of the learning of host MAC addresses in VXLAN or NVGRE network is limited to data plane learning in this document.

http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Table of Contents

1 Introduction

[EVPN] introduces a solution for multipoint L2VPN services, with advanced multi-homing capabilities, using BGP control plane over the core MPLS/IP network. [VXLAN] defines a tunneling scheme to overlay Layer 2 networks on top of Layer 3 networks. [VXLAN] allows for optimal forwarding of Ethernet frames with support for multipathing of unicast and multicast traffic. VXLAN uses UDP/IP encapsulation for tunneling.

In this document, we discuss how Ethernet VPN (EVPN) technology can be used to interconnect VXLAN or NVGRE networks over an MPLS/IP network. This is achieved by terminating the VxLAN tunnel at the hand-off points, performing data plane MAC learning of customer traffic and providing intra-subnet connectivity for the customers at Layer 2 across the MPLS/IP core. The solution maintains control-plane separation among the interconnected VXLAN or NVGRE networks. The scope of the learning of host MAC addresses in VXLAN or NVGRE network is limited to data plane learning in this document. The distribution of MAC addresses in control plane using BGP in VXLAN or NVGRE network is outside of the scope of this document and it is covered in [EVPN-OVERLY].

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

LDP: Label Distribution Protocol. MAC: Media Access Control MPLS: Multi Protocol Label Switching. OAM: Operations, Administration and Maintenance. PE: Provide Edge Node. PW: PseudoWire. TLV: Type, Length, and Value. VPLS: Virtual Private LAN Services. VXLAN: Virtual eXtensible Local Area Network. VTEP: VXLAN Tunnel End Point VNI: VXLAN Network Identifier (or VXLAN Segment ID) ToR: Top of Rack switch. LACP: Link Aggregation Control Protocol

2. Requirements

2.1. Control Plane Separation among VXLAN/NVGRE Networks

It is required to maintain control-plane separation for the underlay networks (e.g., among the various VXLAN/NVGRE networks) being interconnected over the MPLS/IP network. This ensures the following characteristics:

- scalability of the IGP control plane in large deployments and fault domain localization, where link or node failures in one site do not

trigger re-convergence in remote sites.

- scalability of multicast trees as the number of interconnected networks scales.

2.2 All-Active Multi-homing

It is important to allow for all-active multi-homing of the VXLAN/NVGRE network to MPLS/IP network where traffic from a VTEP can arrive at any of the PEs and can be forwarded accordingly over the MPLS/IP network. Furthermore, traffic destined to a VTEP can be received over the MPLS/IP network at any of the PEs connected to the VXLAN/NVGRE network and be forwarded accordingly. The solution MUST support all-active multi-homing to an VXLAN/NVGRE network.

2.3 Layer 2 Extension of VNIs/VSIDs over the MPLS/IP Network

It is required to extend the VXLAN VNIs or NVGRE VSIDs over the MPLS/IP network to provide intra-subnet connectivity between the hosts (e.g. VMs) at Layer 2.

2.4 Support for Integrated Routing and Bridging (IRB)

The data center WAN edge node is required to support integrated routing and bridging in order to accommodate both inter-subnet routing and intra-subnet bridging for a given VNI/VSID. For example, inter-subnet switching is required when a remote host connected to an enterprise IP-VPN site wants to access an application resided on a VM.

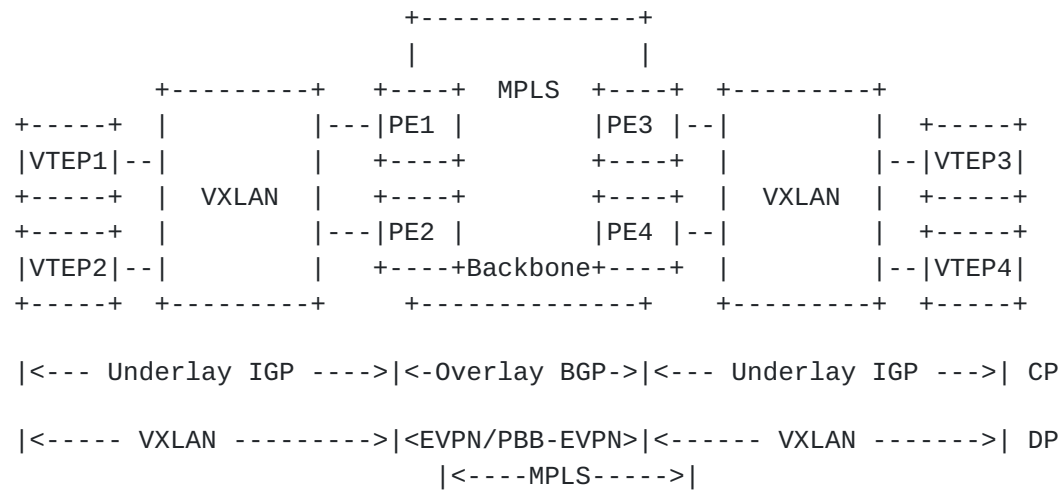
3. Solution Overview

Every VXLAN/NVGRE network, which is connected to the MPLS/IP core, runs an independent instance of the IGP control-plane. Each PE participates in the IGP control plane instance of its VXLAN/NVGRE network.

Each PE node terminates the VXLAN or NVGRE data-plane encapsulation where each VNI or VSID is mapped to a bridge-domain. The PE performs data plane MAC learning on the traffic received from the VXLAN/NVGRE network.

Each PE node implements EVPN or PBB-EVPN to distribute in BGP either the client MAC addresses learnt over the VXLAN tunnel in case of EVPN, or the PEs' B-MAC addresses in case of PBB-EVPN. In the PBB-EVPN case, client MAC addresses will continue to be learnt in data plane.

Each PE node would encapsulate the Ethernet frames with MPLS when sending the packets over the MPLS core and with the VXLAN or NVGRE tunnel header when sending the packets over the VXLAN or NVGRE Network.



Legend: CP = Control Plane View DP = Data Plane View

Figure 1: Interconnecting VXLAN Networks with VXLAN-EVPN

3.1. Redundancy and All-Active Multi-homing

When a VXLAN network is multi-homed to two or more PEs, and provided that these PEs have the same IGP distance to a given NVE, the solution MUST support load-balancing of traffic between the NVE and the MPLS network, among all the multi-homed PEs. This maximizes the use of the bisectional bandwidth of the VXLAN network. One of the main capabilities of EVPN/PBB-EVPN is the support for all-active multi-homing, where the known unicast traffic to/from a multi-homed site can be forwarded by any of the PEs attached to that site. This ensures optimal usage of multiple paths and load balancing. EVPN/PBB-EVPN, through its DF election and split-horizon filtering mechanisms, ensures that no packet duplication or forwarding loops result in such scenarios. In this solution, the VXLAN network is treated as a multi-homed site for the purpose of EVPN operation.

Since the context of this solution is VXLAN networks with data-plane learning paradigm, it is important for the multi-homing mechanism to ensure stability of the MAC forwarding tables at the NVEs, while supporting all-active forwarding at the PEs. For example, in Figure 1 above, if each PE uses a distinct IP address for its VTEP tunnel, then for a given VNI, when an NVE learns a host's MAC address against the originating VTEP source address, its MAC forwarding table will

keep flip-flopping among the VTEP addresses of the local PEs. This is because a flow associated with the same host MAC address can arrive at any of the PE devices. In order to ensure that there is no flip/flopping of MAC-to-VTEP address associations, an IP Anycast address MUST be used as the VTEP address on all PEs multi-homed to a given VXLAN network. The use of IP Anycast address has two advantages:

a) It prevents any flip/flopping in the forwarding tables for the MAC-to-VTEP associations

b) It enables load-balancing via ECMP for DCI traffic among the multi-homed PEs

In the baseline [EVPN] draft, the all-active multi-homing is described for a multi-homed device (MHD) using [LACP] and the single-active multi-homing is described for a multi-homed network (MHN) using [802.1Q]. In this draft, the all-active multi-homing is described for a VXLAN MHN. This implies some changes to the filtering which will be described in details in the multicast section (Section 4.6.2).

The filtering used for BUM traffic of all-active multi-homing in [EVPN] is asymmetric; where the BUM traffic from the MPLS/IP network towards the multi-homed site is filtered on non-DF PE(s) and it passes thorough the DF PE. There is no filtering of BUM traffic originating from the multi-homed site because of the use of Ethernet Link Aggregation: the MHD hashes the BUM traffic to only a single link. However, in this solution because BUM traffic can arrive at both PEs in both core-to-site and site-to-core directions, the filtering needs to be symmetric just like the filtering of BUM traffic for single-active multi-homing (on a per service instance/VLAN basis).

4. EVPN Routes

This solution leverages the same BGP Routes and Attributes defined in [EVPN], adapted as follows:

4.1. BGP MAC Advertisement Route

This route and its associated modes are used to distribute the customer MAC addresses learnt in data plane over the VXLAN tunnel in case of EVPN. Or can be used to distribute the provider Backbone MAC addresses in case of PBB-EVPN.

In case of EVPN, the Ethernet Tag ID of this route is set to zero for VNI-based mode, where there is one-to-one mapping between a VNI and

an EVI. In such case, there is no need to carry the VNI in the MAC advertisement route because BD ID can be derived from the RT associated with this route. However, for VNI-aware bundle mode, where there is multiple VNIs can be mapped to the same EVI, the Ethernet Tag ID MUST be set to the VNI. At the receiving PE, the BD ID is derived from the combination of RT + VNI - e.g., the RT identifies the associated EVI on that PE and the VNI identifies the corresponding BD ID within that EVI.

The Ethernet Tag field can be set to a normalized value that maps to the VNI, in VNI aware bundling services, this would make the VNI value of local significance in multiple Data centers. Data plane need to map to this normalized VNI value and have it on the IP VxLAN packets exchanged between the DCIs.

4.2. Ethernet Auto-Discovery Route

When EVPN is used, the application of this route is as specified in [EVPN]. However, when PBB-EVPN is used, there is no need for this route per [PBB-EVPN].

4.3. Per VPN Route Targets

VXLAN-EVPN uses the same set of route targets defined in [EVPN].

4.4 Inclusive Multicast Route

The EVPN Inclusive Multicast route is used for auto-discovery of PE devices participating in the same tenant virtual network identified by a VNI over the MPLS network. It also enables the stitching of the IP multicast trees, which are local to each VXLAN site, with the Label Switched Multicast (LSM) trees of the MPLS network.

The Inclusive Multicast Route is encoded as follow:

- Ethernet Tag ID is set to zero for VNI-based mode and to VNI for VNI-aware bundle mode.

- Originating Router's IP Address is set to one of the PE's IP addresses.

All other fields are set as defined in [EVPN].

Please see section 4.6 "Handling Multicast"

4.5. Unicast Forwarding

Host MAC addresses will be learnt in data plane from the VXLAN

network and associated with the corresponding VTEP identified by the source IP address. Host MAC addresses will be learnt in control plane if EVPN is implemented over the MPLS/IP core, or in the data-plane if PBB-EVPN is implemented over the MPLS core. When Host MAC addressed are learned in data plane over MPLS/IP core [in case of PBB-EVPN], they are associated with their corresponding BMAC addresses.

L2 Unicast traffic destined to the VXLAN network will be encapsulated with the IP/UDP header and the corresponding customer bridge VNI.

L2 Unicast traffic destined to the MPLS/IP network will be encapsulated with the MPLS label.

4.6. Handling Multicast

Each VXLAN network independently builds its P2MP or MP2MP shared multicast trees. A P2MP or MP2MP tree is built for one or more VNIs local to the VXLAN network.

In the MPLS/IP network, multiple options are available for the delivery of multicast traffic: - Ingress replication - LSM with Inclusive trees - LSM with Aggregate Inclusive trees - LSM with Selective trees - LSM with Aggregate Selective trees

When LSM is used, the trees are P2MP.

The PE nodes are responsible for stitching the IP multicast trees, on the access side, to the ingress replication tunnels or LSM trees in the MPLS/IP core. The stitching must ensure that the following characteristics are maintained at all times:

1. Avoiding Packet Duplication: In the case where the VXLAN network is multi-homed to multiple PE nodes, if all of the PE nodes forward the same multicast frame, then packet duplication would arise. This applies to both multicast traffic from site to core as well as from core to site.

2. Avoiding Forwarding Loops: In the case of VXLAN network multi-homing, the solution must ensure that a multicast frame forwarded by a given PE to the MPLS core is not forwarded back by another PE (in the same VXLAN network) to the VXLAN network of origin. The same applies for traffic in the core to site direction.

The following approach of per-VNI load balancing can guarantee proper stitching that meets the above requirements.

4.6.2. Multicast Stitching with Per-VNI Load Balancing

To setup multicast trees in the VXLAN network for DC applications, PIM Bidir can be of special interest because it reduces the amount of multicast state in the network significantly. Furthermore, it alleviates any special processing for RPF check since PIM Bidir doesn't require any RPF check. The RP for PIM Bidir can be any of the spine nodes. Multiple trees can be built (e.g., one tree rooted per spine node) for efficient load-balancing within the network. All PEs participating in the multi-homing of the VXLAN network join all the trees. Therefore, for a given tree, all PEs receive BUM traffic. DF election procedures of [EVPN] are used to ensure that only traffic to/from a single PE is forwarded, thus avoiding packet duplications and forwarding loops. For load-balancing of BUM traffic, when a PE or an NVE wants to send BUM traffic over the VXLAN network, it selects one of the trees based on its VNI and forwards all the traffic for that VNI on that tree.

Multicast traffic from VXLAN/NVGRE is first subjected to filtering based on DF election procedures of [EVPN] using the VNI as the Ethernet Tag. This is similar to filtering in [EVPN] in principal; however, instead of VLAN ID, VNI is used for filtering, and instead of being 802.1Q frame, it is a VXLAN encapsulated packet. On the DF PE, where the multicast traffic is allowed to be forwarded, the VNI is used to select a bridge domain,. After the packet is de-capsulated, an L2 lookup is performed based on host MAC DA. It should be noted that the MAC learning is performed in data-plane for the traffic received from the VXLAN/NVGRE network and the host MAC SA is learnt against the source VTEP address.

The PE nodes, connected to a multi-homed VXLAN network, perform BGP DF election to decide which PE node is responsible for forwarding multicast traffic associated with a given VNI. A PE would forward multicast traffic for a given VNI only when it is the DF for this VNI. This forwarding rule applies in both the site-to-core as well as core-to-site directions.

4.6.2.1 PIM SM operation

With PIM SM, multicast traffic from the core-to-site could be dropped since a transit router may decide that the RPF path towards the anycast address source is toward a PE node that is not the DF.

The PE nodes whether DF or not, has to forward forward multicast traffic from core-to-side.

The operation would work as follow:

Initially, the PE nodes connected to the multi-homed VXLAN network as well the VTEPs, join towards the RP for the multicast group for a

particular VXLAN.

When BUM traffic needs to be flooded from core to site, all the PE nodes connected to the multi-homed VXLAN network send PIM register messages to the RP. The multicast flow is identified as (anycast address, group) in the register message, and the source address for the PIM-SM register message should be a unique address on the PE node not the anycast address.

The RP will send a join for the (anycast address, group) upon receiving the register message, routed towards the closest PE which could be either the DF or the non-DF. This PE will switch to send traffic natively. Upon receiving the native traffic, the RP will send register-stop messages for other PEs that keep sending registering messages, given that only one PE will get the (anycast address, group) join.

When VTEPs receive traffic from the RP, VTEPs will send (anycast address, group) join, routed towards the closet PE to each VTEP. This starts native forwarding on multiple PE nodes connected to the VXLAN network, but each VTEP or transit router will only accept multicast traffic from one of the multi-homed PE nodes.

If PIM state times out when multicast traffic stops for a period of time, the next flooded packet will trigger the above process again.

It is to be noted that before the RP receives the first natively sent packet from one particular PE node connected to the multihomed VXLAN network, all packets encapsulated in the register messages from all PEs will be forwarded by the RP, causing duplications.

A possible optimization is for all PE nodes connected to the multihomed VXLAN network to send null-register periodically to maintain the PIM state at the RP, instead of encapsulating flooded packets in register messages.

The site-to-core operations for flooding BUM traffic would still be subject to DF election per VNI as described above.

5. NVGRE

Just like VXLAN, all the above specification would apply for NVGRE, replacing the VNI with Virtual Subnet Identifier (VSID) and the VTEP with NVGRE Endpoint.

6. Use Cases Overview

6.1. Homogeneous Network DCI interconnect Use cases

This covers DCI interconnect of two or more VXLAN based Data center over MPLS enabled EVPN core.

6.1.1. VNI Base Mode EVPN Service Use Case

This use case handles the EVPN service where there is one to one mapping between a VNI and an EVI. Ethernet TAG ID of EVPN BGP NLRI should be set to Zero. BD ID can be derived from the RT associated with the EVI/VNI.

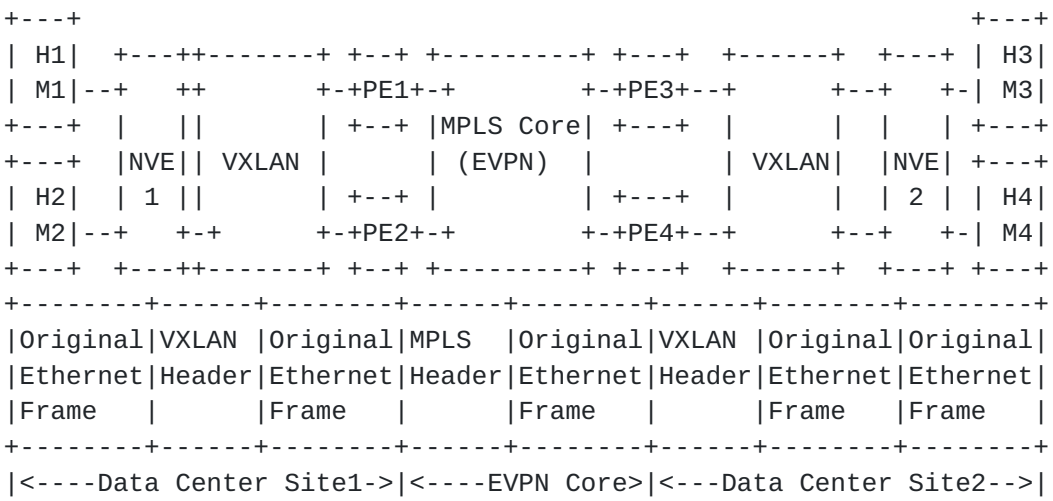


Figure 2 VNI Base Service Packet Flow.

VNI base Service(One VNI mapped to one EVI).

Hosts H1, H2, H3 and H4 are hosts and there associated MAC addresses are M1, M2, M3 and M4. PE1, PE2, PE3 and PE4 are the VXLAN-EVPN gateways. NVE1 and NVE2 are the originators of the VXLAN based network.

When host H1 in Data Center Site1 communicates with H3 in Data Center Site2, H1 forms a layer2 packet with source IP address as IP1 and Source MAC M1, Destination IP as IP3 and Destination MAC as M3(assuming that ARP resolution already happened). VNE1 learns Source MAC and lookup in bridge domain for the Destination MAC. Based on the MAC lookup, the frame needs to be sent to VXLAN network. VXLAN encapsulation is added to the original Ethernet frame and frame is sent over the VXLAN tunnel. Frames arrives at PE1. PE1(i.e. VXLAN gateway), identifies that frame is a VXLAN frame. The VXLAN header is de-capsulated and Destination MAC lookup is done in the bridge domain table of the EVI. Lookup of destination MAC results in the EVPN unicast NH. This NH will be used for identifying the labels (tunnel

label and service label) to be added over the EVPN core. Similar processing is done on the other side of DCI.

6.1.2. VNI Bundle Service Use Case Scenario

In the case of VNI-aware bundle service mode, there are multiple VNIs are mapped to one EVI. The Ethernet TAG ID must be set to the VNI ID in the EVPN BGP NLRIs. MPLS label allocation in this use case scenario can be done either per EVI or per EVI, VNI ID basis. If MPLS label allocation is done per EVI basis, then in data path there is a need to push a VLAN TAG for identifying bridge-domain at egress PE so that Destination MAC address lookup can be done on the bridge domain.

6.1.3. VNI Translation Use Case

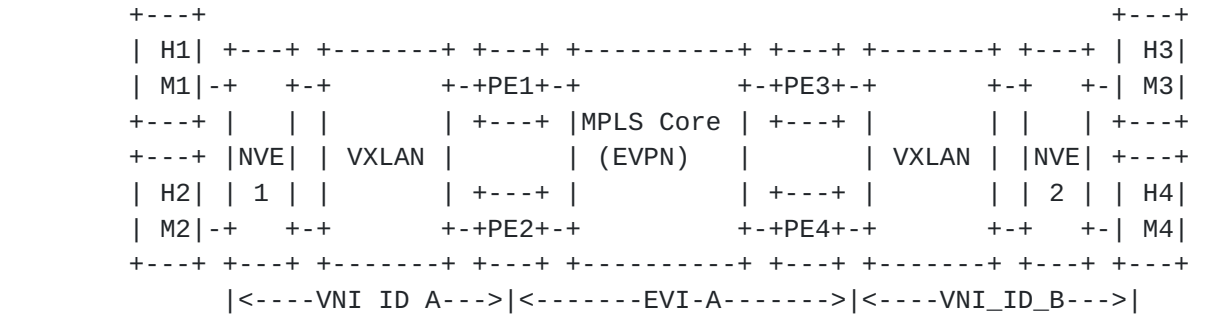


Figure 3 VNI Translation Use Case Scenarios.

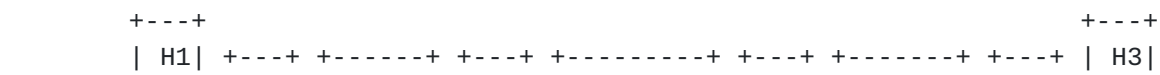
There are two or more Data Center sites. These Data Center sites might use different VNI ID for same service. For example, Service A usage "VNI_ID_A" at data center site1 and "VNI_ID_B" for same service in data center site 2. VNI ID A is terminated at ingress EVPN PE and VNI ID B is encapsulated at the egress EVPN PE.

6.2. Heterogeneous Network DCI Use Cases Scenarios

Data Center sites are upgraded slowly; so heterogeneous network DCI solution is required from the perspective of migration approach from traditional data center to VXLAN based data center. For Example Data Center Site1 is upgrade to VXLAN but Data Center Site 2 and 3 are still layer2/VLAN based data centers. For these use cases, it is required to provide VXLAN VLAN interworking over EVPN core.

6.2.1. VXLAN VLAN Interworking Over EVPN Use Case Scenario

The new data center site is VXLAN based data center site. But the older data center sites are still based on the VLAN.



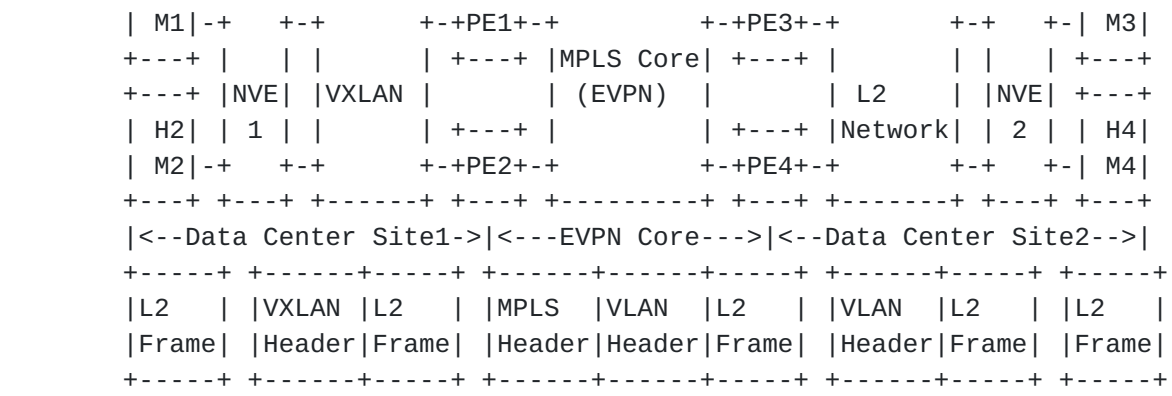


Figure 5 VXLAN VLAN interworking over EVPN Use Case.

If a service that are represented by VXLAN on one site of data center and via VLAN at different data center sites, then it is a recommended to model the service as a VNI base EVPN service. The BGP NLRIs will always advertise VLAN ID TAG as '0' in BGP routes. The advantage with this approach is that there is no requirement to do the VNI normalization at EVPN core. VNI ID A is terminated at ingress EVPN PE and "VLAN ID B" is encapsulated at the egress EVPN PE.

7. Acknowledgements

The authors would like to acknowledge Wen Lin contributions to this document.

8. Security Considerations

There are no additional security aspects that need to be discussed here.

9. IANA Considerations

10. References

10.1 Normative References

[KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

10.2 Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", RFC 7432, February, 2012.

[PBB-EVPN] Sajassi et al., "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, September, 2015.

[VXLAN] Mahalingam, Dutt et al., A Framework for Overlaying

Virtualized Layer 2 Networks over Layer 3 Networks, RFC 7348, August, 2012.

[NVGRE] Sridharan et al., Network Virtualization using Generic Routing Encapsulation, RFC 7637, July, 2012.

Authors' Addresses

Sami Boutros
VMware, Inc.
EMail: sboutros@vmware.com

Ali Sajassi
Cisco Systems
EMail: sajassi@cisco.com

Samer Salam
Cisco Systems
EMail: ssalam@cisco.com

Dennis Cai
Cisco Systems
EMail: dcai@cisco.com

Tapraj Singh
Juniper Networks
Email: tsingh@juniper.net

John Drake
Juniper Networks
Email: jdrake@juniper.net

Samir Thoria
Cisco
EMail: sthoria@cisco.com

Jeff Tantsura
Ericsson
Email: jeff.tantsura@ericsson.com