INTERNET-DRAFT
Intended Status: Informational

Sami Boutros
Ali Sajassi
Samer Salam
Dennis Cai

Expires: August 28, 2013
February 24, 2013

VXLAN DCI Using EVPN
draft-boutros-l2vpn-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.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

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

Copyright and License Notice

Table of Contents

1 Introduction

[E-VPN] 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 (E-VPN) 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 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.

2. Requirements

2.1. Control Plane Separation among VXLAN/NVGRE Networks

It is required to maintain control-plane separation 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 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.

2.3 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.

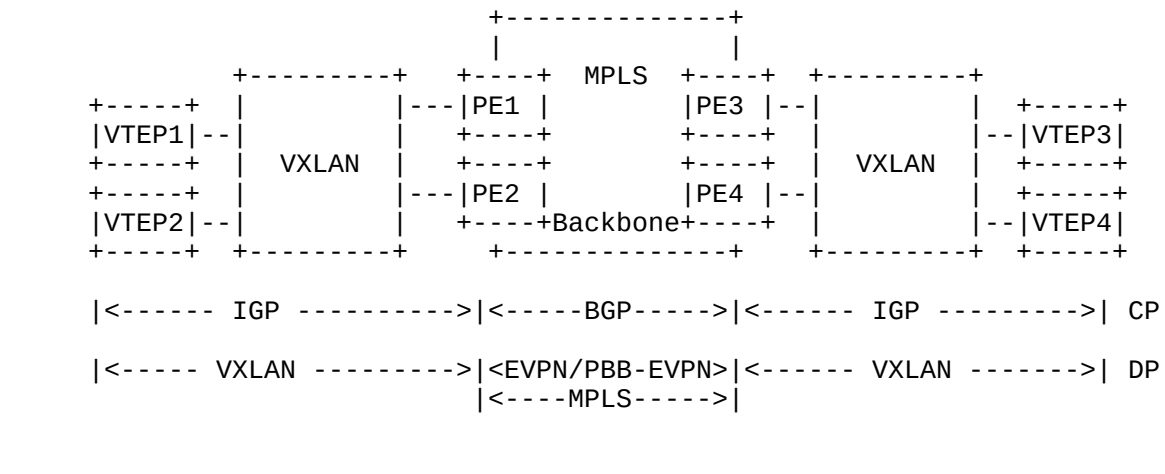
3. Solution Overview

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 local site.

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 E-VPN or PBB-EVPN to distribute 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

4. E-VPN Routes This solution leverages the same BGP Routes and Attributes defined in [E-VPN], 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.

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 [E-VPN].

4.4 Inclusive Multicast Route

The E-VPN Inclusive Multicast route is used to distribute the VNI information over the MPLS network. This is required to perform the discovery of the PEs participating in a given VNI. 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 VXLAN Network Identifier (VNI).

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

All other fields are set as defined in [E-VPN].

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. Host MAC addresses will be learnt in control plane if E-VPN 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

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.

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. Acknowledgements

TBD.

7. Security Considerations

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

8. IANA Considerations

TBD.

9. References

9.1 Normative References

[KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2 Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-00.txt, work in progress, February, 2012.

[TRILL] Sajassi et al., TRILL-EVPN draft-ietf-l2vpn-trill-evpn-00, work in progress, June 2012.

[VXLAN] Mahalingam, Dutt et al., A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks draft-mahalingam-dutt-dcops-vxlan-02.txt, work in progress, August, 2012.

[NVGRE] Sridharan et al., Network Virtualization using Generic Routing Encapsulation draft-sridharan-virtualization-nvgre-01.txt, work in progress, July, 2012.

Authors' Addresses

Sami Boutros
Cisco Systems

EMail: sboutros@cisco.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