INTERNET-DRAFT
Intended Status: Informational

Expires: August 22, 2013

Sami Boutros
Ali Sajassi
Samer Salam
Dennis Cai

February 18, 2013

VXLAN-EVPN
draft-boutros-l2vpn-vxlan-evpn-00.txt

Abstract

This document describes how Ethernet VPN (E-VPN) technology can be used to interconnect VXLAN and NVGRE networks over an MPLS/IP network, while maintaining C-MAC address transparency on the hand-off point and control-plane isolation among the interconnected VXLAN and NVGRE networks.

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 for distributing customer/client MAC (C-MAC) address reach-ability information over the core MPLS/IP network. [VXLAN] defines a a tunneling scheme to overlay Layer 2 networks on top of Layer 3 networks. Similar to [TRILL], [VXLAN] allows for optimal forwarding of Ethernet frames with support for multipathing of unicast and multicast traffic. VXLAN uses an IP/UDP encapsulation for tunneling. In this document, we discuss how Ethernet VPN (E-VPN) technology can be used to interconnect VXLAN networks over an MPLS/IP network, while guaranteeing C-MAC address transparency on the hand-off point and control-plane isolation among the interconnected VXLAN networks.

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. C-MAC Address Transparency on the Hand-off Point

When VXLAN networks are interconnected over an MPLS/IP network, it is required to maintain C-MAC address transparency on the hand-off point and the edge (i.e. PE) of the MPLS network. Otherwise, the MPLS edge

nodes may suffer from MAC address table space exhaustion, as they would need to learn the C-MAC addresses from all interconnected VXLAN networks.

VXLAN-EVPN supports seamless interconnect with VXLAN while guaranteeing C-MAC address transparency on the PE nodes.

2.2. Control Plane Isolation among VXLAN Networks

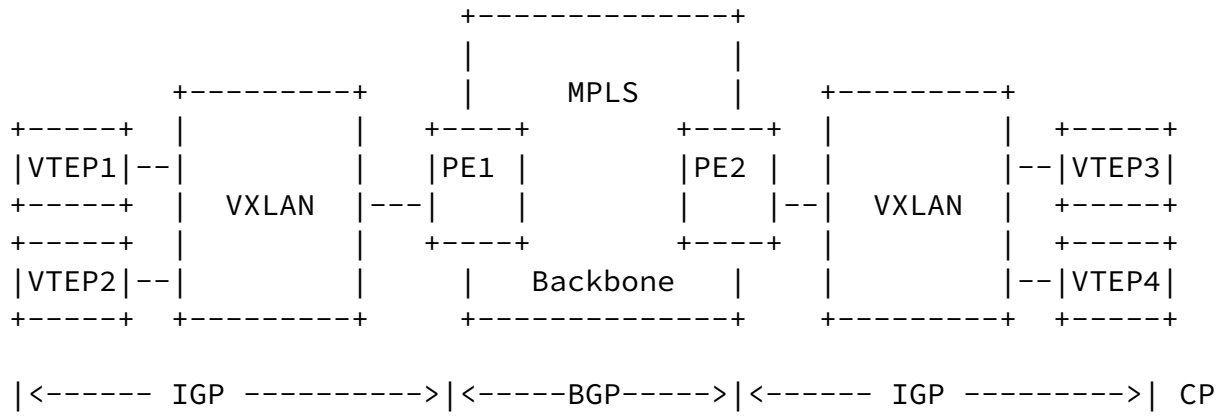
It is required to maintain control-plane isolation among the various VXLAN networks being interconnected over the MPLS/IP network. This ensures the following characteristics:

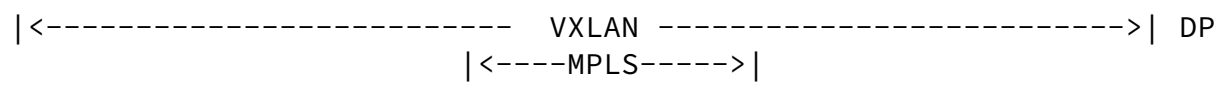
- scalability of the IGP control plane in large deployments and fault domain localization, where link or node failures in one site do not trigger reconvergence in remote sites.

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

3. Solution Overview

Every VXLAN network that is connected to the MPLS core runs an independent instance of the IGP control-plane. Each PE participates in the IGP control plane instance of its local site. The PE does not terminate the VXLAN data-plane encapsulation. The PE nodes encapsulate VXLAN IP/UDP packets with MPLS in the imposition path, and de-capsulate them in the disposition path.





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

Figure 1: Interconnecting VXLAN Networks with VXLAN-EVPN

4. EVPN Routes

VXLAN-EVPN 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 not needed in VXLAN-EVPN.

4.2. Ethernet Auto-Discovery Route

This route and its associated modes are not needed in VXLAN-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

TBD

4.6. Handling Multicast

Each VXLAN network independently builds its P2MP or MP2MP shared multicast trees. A P2MP or MP2MP tree is built for every VNI 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 may be either P2MP or MP2MP.

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.

3. Pacifying RPF Checks: For multicast traffic originating from a

different VXLAN network, the RPF check shouldn't have an issue with PIM-BIDIR.

There are two approaches by which the above operation can be guaranteed: one offers per-VTEP source load-balancing while the other offers per- VNI load-balancing.

4.6.1. Multicast Stitching with Per-Source Load Balancing

The PE nodes, connected to a multi-homed VXLAN network, perform BGP DF election to decide which PE is responsible for forwarding multicast traffic from a given source VTEP. A PE would only forward multicast traffic from source VTEP for which it is the DF, in both the site to core as well as core to site directions. This solves both the issue of avoiding packet duplication as well as the issue of avoiding forwarding loops.

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, NVGRE IP tunnels encapsulating Customer Ethernet Frames can be carried seamlessly over EVPN, 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

EMail: sajassi@cisco.com

Samer Salam
Cisco Systems

EMail: ssalam@cisco.com

Dennis Cai
Cisco Systems

EMail: dcai@cisco.com