BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2019

Ali. Sajassi
Mankamana. Mishra
Samir. Thoria
Patrice. Brissette
Cisco Systems

October 22, 2018

AC-Aware Bundling Service Interface in EVPN
draft-sajassi-bess-evpn-ac-aware-bundling-00

Abstract

EVPN provides an extensible and flexible multi-homing VPN solution over an MPLS/IP network for intra-subnet connectivity among Tenant Systems and End Devices that can be physical or virtual.

EVPN multihoming with IRB is one of the common deployment scenarios. There are deployments which requires capability to have multiple subnets designated with multiple VLAN IDs in single bridge domain.

RFC7432 defines three different type of service interface which serve different requirements but none of them address the requirement to be able to support multiple subnets within single bridge domain. In this draft we define new service interface type to support multiple subnets in single bridge domain.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on April 25, 2019.

Table of Contents

1. Introduction

EVPN based multi-homing is becoming the basic building block for providing redundancy in next generation data center deployments as well as service provider access/aggregation network. For EVPN IRB mode, there are deployments which expect to be able to support multiple subnets within single Bridge Domain. Each subnets would be differentiated by VLAN. Thus, single IRB interface can still serve multiple subnets.

Motivation behind such deployments are

1. Manageability: If there is support to have multiple subnets using single bridge domain, it would require only one Bridge domain and one IRB for "N" subnets compare to "N" Bridge domain and "N" IRB interfaces to manage.

2. Simplicity: It avoids extra configuration by configuring Vlan Range as compare to individual VLAN, BD and IRB interface per subnet.

Multiple subnet per bridge domain deployments require that there would not be duplicate MAC address across subnet.

[RFC7432] defines three types of service interfaces. None of them provide flexibility to achieve multiple subnet within single bridge domain. Brief about existing service interface from [RFC7432] are ,

1. VLAN-Based Service Interface: With this service interface, an EVPN instance consists of only a single broadcast domain (e.g., a single VLAN). Therefore, there is a one-to-one mapping between a VID on this interface and a MAC-VRF.

2. VLAN Bundle Service Interface: With this service interface, an EVPN instance corresponds to multiple broadcast domains (e.g., multiple VLANs); however, only a single bridge table is maintained per MAC-VRF, which means multiple VLANs share the same bridge table. The MPLS-encapsulated frames MUST remain tagged with the originating VID. Tag translation is NOT permitted. The Ethernet Tag ID in all EVPN routes MUST be set to 0.

3. VLAN-Aware Bundle Service Interface: With this service interface, an EVPN instance consists of multiple broadcast domains (e.g., multiple VLANs) with each VLAN having its own bridge table -- i.e., multiple bridge tables (one per VLAN) are maintained by a single MAC-VRF corresponding to the EVPN instance.

Though from definition it looks like VLAN Bundle Service Interface does provide flexibility to support multiple subnet within single bridge domain. But it can not serve the requirement which is being described in this draft. For example, lets take the case from Figure-1, If PE1 learns MAC of H1 on Vlan 1 (subnet S1). When MAC route is originated , as per [RFC7432] ether tag would be set to 0. If there is packet coming from IRB interface which is untagged packet, and it reaches to PE2, PE2 does not have associated AC information. In this case PE2 can not forward traffic which is destined to H1.

This draft proposes an extension to existing service interface types defined in [RFC7432] and defines AC-aware Bundling service interface. AC-aware Bundling service interface would provide mechanism to have multiple subnets in single bridge domain. This extension is applicable only for multi-homed EVPN peers.

With this proposal IRB interface could either have multiple subnets or an aggregate subnet representing all individual subnets (when such aggregation is possible).

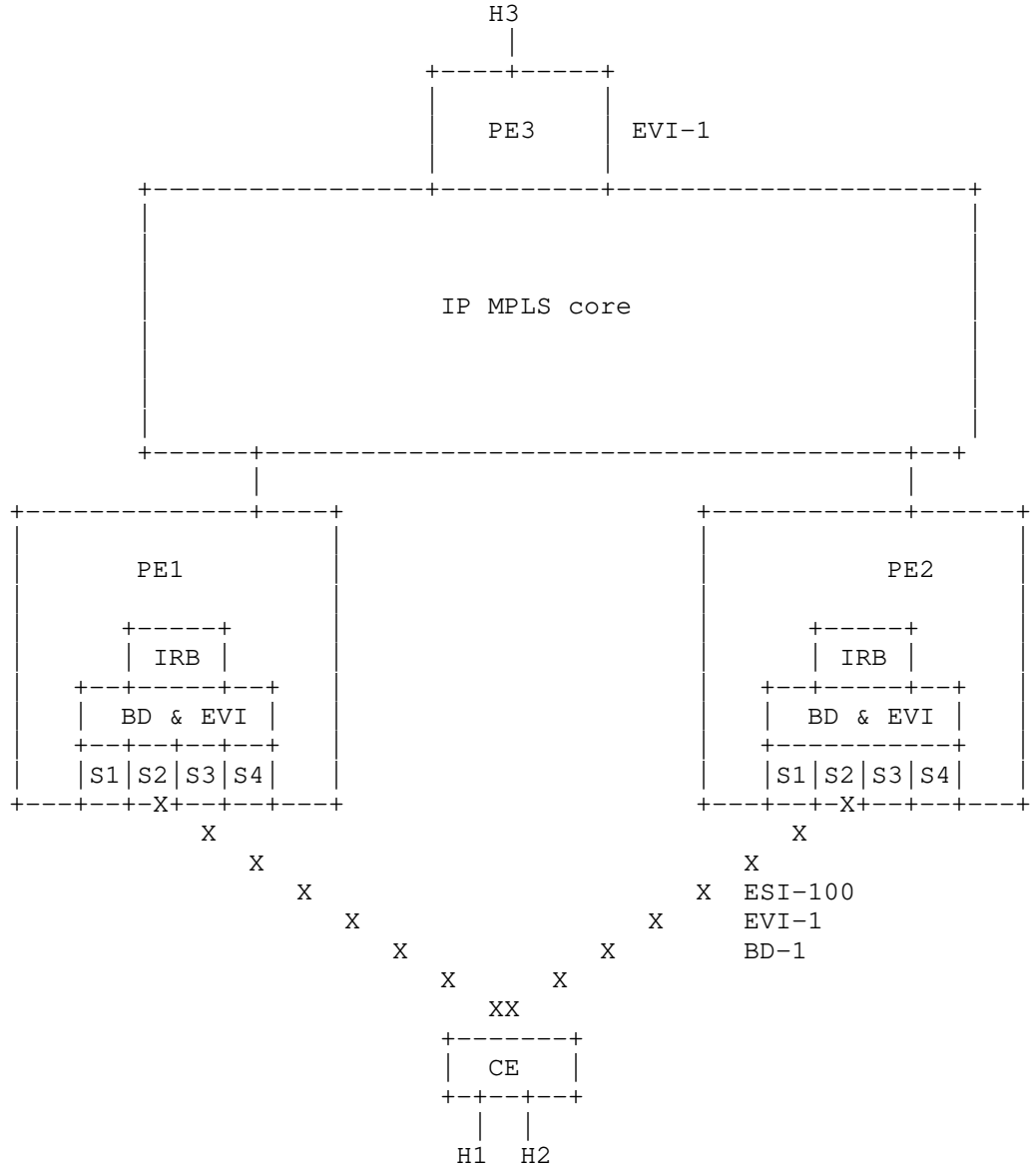

Figure 1: EVPN topology with multi-homing and non multihoming peer

The above figure shows sample EVPN topology, PE1 and PE2 are multihomed peers. PE3 is remote peer which is part of same EVPN instance (evi1). It is showing four subnets S1, S2, S3, S4 where numeric value provides associated Vlan information.

1.1. Problem with Unicast MAC route processing for multihome case

BD-1 has multiple subnets where each subnet is distinguished by Vlan 1, 2 ,3 and 4. PE1 learns MAC address MAC-1 from AC associated with subnet S1. PE1 uses MAC route to advertise MAC-1 presence to peer PEs. As per [RFC7432] MAC route advertisement from PE1 does not carry any context which can provide information about MAC address association with AC. When PE2 receives MAC route with MAC-2 it can not determine which AC this MAC belongs too.

Since PE2 could not bind MAC-1 with correct AC, when it receives data traffic destined to MAC-1, it can not find correct AC where data MUST be forwarded.

1.2. Problem with Multicast route synchronization

[I-D.ietf-bess-evpn-igmp-mld-proxy] defines mechanism to synchronize multicast routes between multihome peer. In above case if Receiver behind S1 send IGMP membership request, CE could hash it to either of the PE. When Multicast route is originated, it does not contain any AC information. Once it reaches to remote PE, it does not have any information about which subnet this IGMP membership request belong to.

1.3. Potential Security concern caused by misconfiguration

In case of single subnet per bridge domain, there is potential case of security issue. For example if PE1 , BD1 is configured with Vlan-1 where as multihome peer PE2 has configured Vlan-2. Now each of the IGMP membership request on PE1 would be synchronized to PE2. and PE2 would process multicast routes and start forwarding multicast traffic on Vlan-2, which was not intended.

2. 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 [RFC2119] .

AC: Attachment Circuit.

ARP: Address Resolution Protocol.

BD: Broadcast Domain. As per [RFC7432], an EVI consists of a single or multiple BDs. In case of VLAN-bundle and VLAN-based service models (see [RFC7432]), a BD is equivalent to an EVI. In case of VLAN-aware bundle service model, an EVI contains multiple BDs. Also, in this document, BD and subnet are equivalent terms.

BD Route Target: refers to the Broadcast Domain assigned Route Target [RFC4364]. In case of VLAN-aware bundle service model, all the BD instances in the MAC-VRF share the same Route Target.

BT: Bridge Table. The instantiation of a BD in a MAC-VRF, as per [RFC7432].

DGW: Data Center Gateway.

Ethernet A-D route: Ethernet Auto-Discovery (A-D) route, as per [RFC7432].

Ethernet NVO tunnel: refers to Network Virtualization Overlay tunnels with Ethernet payload. Examples of this type of tunnels are VXLAN or GENEVE.

EVI: EVPN Instance spanning the NVE/PE devices that are participating on that EVPN, as per [RFC7432].

EVPN: Ethernet Virtual Private Networks, as per [RFC7432].

GRE: Generic Routing Encapsulation.

GW IP: Gateway IP Address.

IPL: IP Prefix Length.

IP NVO tunnel: it refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload)

IP-VRF: A VPN Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by EVPN and IP-VPN address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.

IRB: Integrated Routing and Bridging interface. It connects an IP-VRF to a BD (or subnet).

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on an NVE/PE, as per [RFC7432]. A MAC-VRF is also an instantiation of an EVI in an NVE/PE.

ML: MAC address length.

ND: Neighbor Discovery Protocol.

NVE: Network Virtualization Edge.

GENEVE: Generic Network Virtualization Encapsulation, [GENEVE].

NVO: Network Virtualization Overlays.

RT-2: EVPN route type 2, i.e., MAC/IP advertisement route, as defined in [RFC7432].

RT-5: EVPN route type 5, i.e., IP Prefix route. As defined in Section 3 of [EVPN-PREFIX].

SBD: Supplementary Broadcast Domain. A BD that does not have any ACs, only IRB interfaces, and it is used to provide connectivity among all the IP-VRFs of the tenant. The SBD is only required in IP-VRF- to-IP-VRF use-cases (see Section 4.4.).

SN: Subnet.

TS: Tenant System.

VA: Virtual Appliance.

VNI: Virtual Network Identifier. As in [RFC8365], the term is used as a representation of a 24-bit NVO instance identifier, with the understanding that VNI will refer to a VXLAN Network Identifier in VXLAN, or Virtual Network Identifier in GENEVE, etc. unless it is stated otherwise.

VTEP: VXLAN Termination End Point, as in [RFC7348].

VXLAN: Virtual Extensible LAN, as in [RFC7348].

This document also assumes familiarity with the terminology of [RFC7432],[RFC8365], [RFC7365].

3. Requirements

1. Service interface MUST be able to support multiple subnets designated by Vlan under single bridge domain.

2. New Service interface handling procedure MUST make sure to have backward compatibility with implementation procedures defined in [RFC7432]

3. New Service interface MUST be extendible to multicast routes defined in [I-D.ietf-bess-evpn-igmp-mld-proxy] too.

4. Solution Description

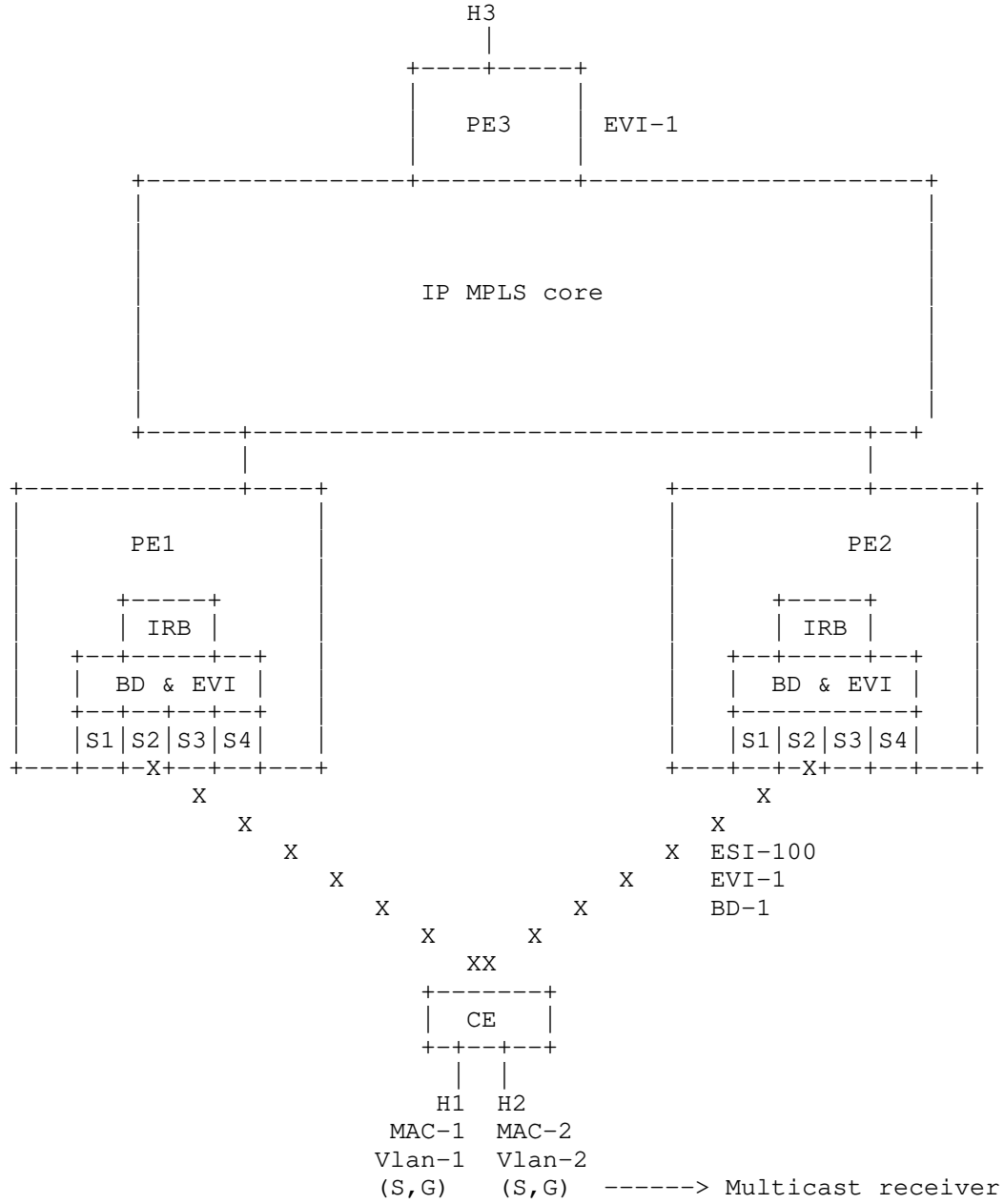

Figure 2: AC aware bundling procedures

Consider the above topology, where AC aware bundling service interface is supported. Host H1 on Vlan-1 has MAC address as MAC-1 and Host H2 on Vlan 2 has MAC address as MAC-2.

4.1. Control Plane Operation

4.1.1. MAC/IP Address Advertisement

4.1.1.1. Local Unicast MAC learning

1. [RFC7432] section 9.1 describes different mechanism to learn Unicast MAC address locally. PEs where AC aware bundling is supported, MAC address is learnt along with Vlan associated with AC.

2. MAC/IP route construction follows mechanism defined in [RFC7432] section 9.2.1. Along with RT-2 it must attach Attachment Circuit ID Extended Community (Section 5.1).

3. From Figure-2 PE1 learns MAC-1 on S1. It MUST construct MAC route with procedure defined in [RFC7432] section 9.2.1. It MUST attach Attachment Circuit ID Extended Community (Section 5.1).

4.1.1.2. Remote Unicast MAC learning

1. Presence of Attachment Circuit ID Extended Community (Section 5.1) MUST be ignored by non multihoming PEs. Remote PE (Non Multihome PE) MUST process MAC route as defined in [RFC7432]

2. Multihoming peer MUST process Attachment Circuit ID Extended Community (Section 5.1) to attach remote MAC address to appropriate AC.

3. From Figure-2 PE3 receives MAC route for MAC-1. It MUST ignore AC information in Attachment Circuit ID Extended Community (Section 5.1) which was received with RT-2.

4. PE2 receives MAC route for MAC-1. It MUST get Attachment Circuit ID from Attachment Circuit ID Extended Community (Section 5.1) in RT-2 and associate MAC address with specific subnet.

4.1.2. Multicast route Advertisement

4.1.2.1. Local multicast state

When a local multihomed bridge port in given BD receives IGMP membership request and ES is operating in All-active or Single-Active redundancy mode, it MUST synchronize multicast state by originating

multicast route defined in section 7 of [I-D.ietf-bess-evpn-igmp-mld-proxy]. When Service interface is AC aware it MUST attach Attachment Circuit ID Extended Community (Section 5.1) along with multicast route. For example in Figure-2 when H2 sends IGMP membership request for (S,G) , CE hashed it to one of the PE. Lets say PE1 received IGMP membership request, now PE1 MUST originate multicast route to synchronize multicast state with PE2. Multicast route MUST contain Attachment Circuit ID Extended Community (Section 5.1) along with multicast route.

If PE1 had already originated multicast route for (S,G) from subnet S2. Now if host H1 also sends IGMP membership request for (S,G) on subnet S1, PE1 MUST originate route update with Attachment Circuit ID Extended Community (Section 5.1).

4.1.2.2. Remote multicast state

If multihomed PE receives remote multicast route on Bridge Domain for given ES, route MUST be programmed to correct subnet. Subnet information MUST be get from Attachment Circuit ID Extended Community. For example PE2 receives multicast route on Bridge Domain BD-1 for ES ESI-100, From Attachment Circuit ID Extended Community (Section 5.1) it receives AC information and associates multicast route (S,G) to subnet S2.

When PE2 receives route update with Attachment Circuit ID Extended Community added for subnet S1, port associated with subnet S1 MUST be added for multicast route.

4.2. Data Plane Operation

4.2.1. Unicast Forwarding

1. Packet received from CE must follow same procedure as defined in [RFC7432] section 13.1

2. Unknown Unicast packets from a Remote PE MUST follow procedure as per [RFC7432] section 13.2.1.

3. Known unicast Received on a Remote PE MUST follow procedure as per [RFC7432] section 13.2.2. So in Figure-2 if PE3 receives known unicast packet for destination MAC MAC-1, it MUST follow procedure defined in [RFC7432] section 13.2.2.

4. If destination MAC lookup is performed on known unicast packet, destination MAC lookup MUST provide Vlan and Port tuple. For example if PE2 receives unicast packet which is destined to MAC-1 (packet might be coming from IRB or remote PE with EVPN tunnel),

destination MAC lookup on PE2 MUST provide outgoing port along with associated MAC address. In this case traffic MUST be forwarded to S1 with Vlan 1.

4.2.2. Multicast Forwarding

1. Multicast traffic from CE and remote PE MUST follow procedure defined in [RFC7432]

2. When multicast traffic is being received on IRB Interface, layer-3 forwarding is based on traditional multicast without any new modification. On bridge domain multicast traffic is forwarded towards right AC based on multicast state.

5. BGP Encoding

This document defines one new BGP Extended Community for EVPN.

5.1. Attachment Circuit ID Extended Community

A new EVPN BGP Extended Community called Attachment Circuit ID is introduced here. This new extended community is a transitive extended community with the Type field of 0x06 (EVPN) and the Sub-Type of TBD. It is advertised along with EVPN MAC/IP Advertisement Route (Route Type 2) per [RFC7432] for AC-Aware Bundling Service Interface.

The Attachment Circuit ID Extended Community is encoded as an 8-octet value as follows:

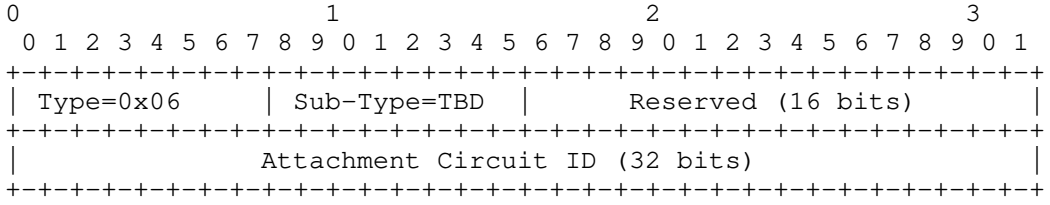

Figure 3: Attachment Circuit ID Extended Community

This extended community is used to carry the Attachment Circuit ID associated with the received MAC address and it is advertised along with EVPN MAC/IP and EVPN multicast Advertisement route. The receiving PE who is a member of an All-Active or Single-Active multi-homing group uses this information to not only synchronize the MAC address but also the associated AC over which the MAC addresses is received.

6. Security Considerations

The same Security Considerations described in [RFC7432] are valid for this document.

7. IANA Considerations

A new transitive extended community Type of 0x06 and Sub-Type of TBD for EVPN Attachment Circuit Extended Community needs to be allocated by IANA.

8. Acknowledgement

9. References

9.1. Normative References

[I-D.ietf-bess-evpn-igmp-mld-proxy]
Sajassi, A., Thoria, S., Patel, K., Yeung, D., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-02 (work in progress), June 2018.

[I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10 (work in progress), August 2018.

9.2. Informative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, DOI 10.17487/RFC7365, October 2014, <https://www.rfc-editor.org/info/rfc7365>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

Authors' Addresses

Ali Sajassi
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: sajassi@cisco.com

Mankamana Mishra
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: mankamis@cisco.com

Samir Thoria
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: sthoria@cisco.com

Patrice Brissette
Cisco Systems

Email: pbrisset@cisco.com