BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: 16 March 2023

A. Sajassi
P. Brissette
M. Mishra
S. Thoria
Cisco Systems
J. Rabadan
Nokia
J. Drake
Juniper Networks
12 September 2022

AC-Aware Bundling Service Interface in EVPN
draft-sajassi-bess-evpn-ac-aware-bundling-06

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 Broadcast Domain.

EVPN technology defines three different types of service interface which serve different requirements but none of them address the requirement of supporting multiple subnets within single Broadcast Domain. In this draft we define new service interface type to support multiple subnets in single Broadcast Domain. Service interface proposed in this draft will be applicable to multihoming case only.

Requirements Language

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] and RFC 8174 [RFC8174].

Status of This Memo

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

Copyright Notice

Table of Contents

1. Introduction

EVPN based All-Active 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 Broadcast Domain. Each subnet would be differentiated by VLAN. Thus, single IRB interface can still serve multiple subnet.

Motivation behind such deployments are

1. Manageability: The support to have multiple subnets using single Broadcast Domain requires only one Broadcast Domain and one IRB for "N" subnets compare to "N" Broadcast Domain and "N" IRB interface to manage.

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

[RFC7432] defines three types of service interface. None of them provide flexibility to achieve multiple subnets within single Broadcast Domain. The different types of 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.

From definition, it seems like VLAN Bundle Service Interface does provide flexibility to support multiple subnets within single Broadcast Domain. However, the requirement is to have multiple subnets from same ES on multi-homing all active mode; that would not work. For example, lets take the case from Figure 1 where PE1 learns MAC of H1 on VLAN 1 (subnet S1). PE1 originates EVPN MAC route, as per [RFC7432], where the Ethernet Tag would be set to 0. Incoming packets from IRB interface, at PE2, are untagged packet. PE2 does not have any associated AC information from EVPN MAC routes advertised by PE1. 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 Broadcast Domain. This extension is applicable only for multi-homed EVPN peers.

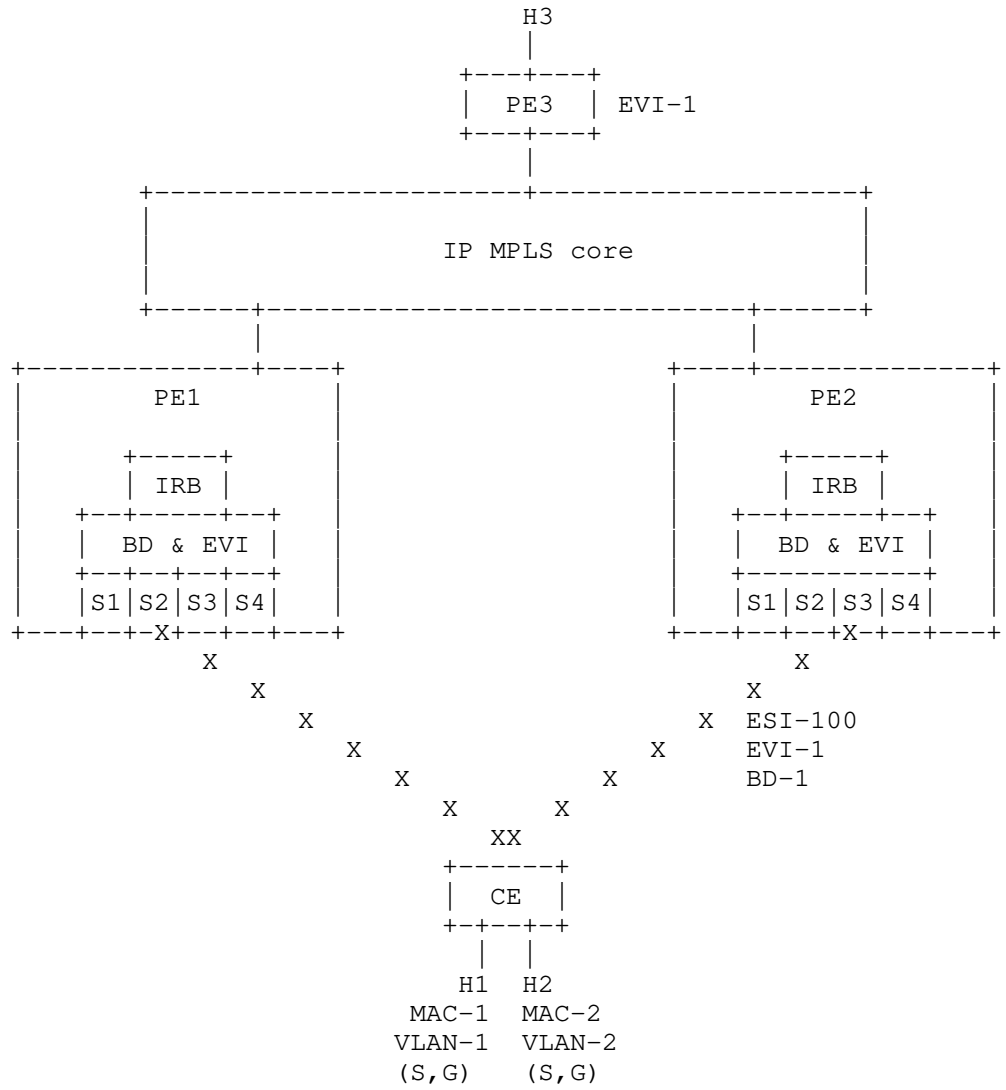


Figure 1

EVPN topology with multi-homing and non multihoming peer.

Figure 1 shows sample EVPN topology where PE1 and PE2 are multihomed peers. PE3 is remote peer participating in the same EVPN instance (EVI-1). It illustrates four subnets S1, S2, S3 and S4 where numerical value provides associated VLAN information.

1.1. Problem With Unicast MAC Route

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 providing 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 does not know the destination AC since multiple bridge ports have the same ESI assignment.

1.2. Problem With Multicast Route Synchronization

[RFC9251] defines mechanism to synchronize multicast routes between multihome peers. In above case, if receiver behind S1 send IGMP membership request, CE could hash it to either of the PEs. When multicast route is originated, it does not contain any AC information. Once it reaches to peering PE, it does not have any information about which subnet this IGMP membership request belong to. Similarly to unicast traffic problem, the incoming multicast traffic from IRB cannot be forearded to proper AC.

1.3. Potential Security Concern caused By Misconfiguration

In case of single subnet per Broadcast Domain, there is potential case of security issue. For example, PE1 has BD1 configured with VLAN-1 where as multihome peer PE2 has BD1 configured VLAN-2. Each of the IGMP membership requests 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. Again, similar issue can potentially be seen with unicast traffic.

2. Terminology

* 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].

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

* 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].

* 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.

* ND: Neighbor Discovery Protocol.

* RD: BGP Route Distinguisher.

* 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 [RFC9136].

* SN: Subnet.

* TS: Tenant System.

* VLAN: The usage of VLAN refers to 802.1Q or 802.1AD tag.

* (S,G): Multicast membership request

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

3. Requirements

1. A service interface represents an attachement-circuit where multiple VLAN are configured on. Each of these VLANs are represented by a different AC under a single Broadcast Domain.

2. Single Broadcast Domain MUST support service interfaces.

3. Service interface MUST be applicable to multihomed peers only.

4. Service interface MUST have an Ethernet-Segment identifier assignement.

5. New service interface handling procedures MUST be backward compatible with implementation procedures defined in [RFC7432]

6. New service interface MUST support EVPN multicast routes defined in [RFC9251] too.

4. Solution Description

4.1. Control Plane Operation

4.1.1. MAC/IP Address Advertisement

4.1.1.1. Local Unicast MAC Learning

[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.

MAC/IP route construction follows mechanism defined in [RFC7432] section 9.2.1. An attach Attachment Circuit ID Extended Community (Section 6.1) must be attached to EVPN RT-2.

4.1.1.2. Remote Unicast MAC Learning

Presence of Attachment Circuit ID Extended Community (Section 6.1) MUST be ignored by non multihoming PEs. Remote PE (non-multihome PE) MUST process MAC route as defined in [RFC7432]

Multihoming peer MUST process Attachment Circuit ID Extended Community (Section 6.1) to attach remote MAC address to appropriate AC.

From Figure 1, PE2 receives MAC route for MAC-1. It MUST get Attachment Circuit ID from Attachment Circuit ID Extended Community (Section 6.1) 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 AC in given Broadcast Domain receives IGMP membership request, it MUST synchronize multicast state by originating multicast route defined in [RFC9251]. When Service interface is AC aware it MUST attach Attachment Circuit ID Extended Community (Section 6.1) along with multicast route. For example in Figure 1 when H2 sends IGMP membership request for (S,G), CE hashed it to one of the PE. Lets say PE1 received IGMP membership request. PE1 MUST originate multicast route to synchronize multicast state with PE2. Multicast route MUST contain Attachment Circuit ID Extended Community (Section 6.1) along with multicast route.

PE1 must originate multicast route updates for any subsequent IGMP membership requests under same or different subnet attaching adequate Attachment Circuit ID Extended Community (Section 6.1).

4.1.2.2. Remote Multicast State

If multihomed PE receives remote multicast route on Broadcast Domain for given ES, route MUST be programmed to correct subnet. Subnet information MUST be extracted from Attachment Circuit ID Extended Community. That value maps to the VLAN of a local AC where the multicast route is associated to.

4.2. Data Plane Operation

4.2.1. Unicast Forwarding

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

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

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

If destination MAC lookup is performed on known unicast packet, destination MAC lookup MUST provide VLAN and local AC information. 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 VLAN value.

4.2.2. Multicast Forwarding

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

Multicast traffic received from IRB interface or EVPN tunnel, route lookup would be performed based on IGMP snooping state and traffic would be forwarded to appropriate AC.

5. Mis-configuration Across Multihoming Peers

If there is mis-configuration of VLAN or VLAN range across multihoming peers, same MAC address would be learnt with different VLAN per Broadcast Domain. In this case Error message MUST be thrown for operator to make configuration changes. Furthermore, the errored MAC route MUST be ignored.

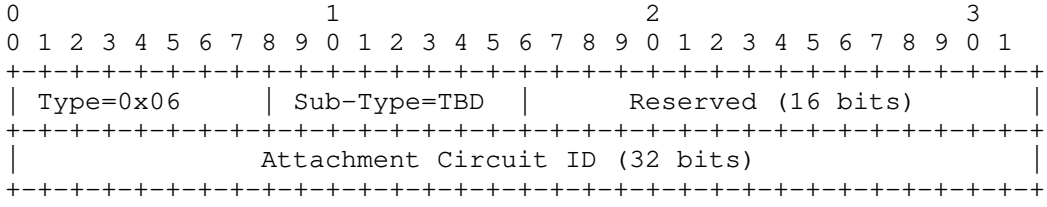
6. BGP Encoding

This document defines one new BGP Extended Community for EVPN.

6.1. Attachment Circuit ID Extended Community

A new EVPN BGP Extended Community called Attachment Circuit ID is introduced. 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. It may also be advertised along with EVPN Multicast Route (Route Type 7 and 8) as per [RFC9251]. Generically speaking, the new extended community must be attached to any routes which require specific VLAN identification.

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



Attachment Circuit ID Extended Community

The attachment circuit ID plays the role of normalized VID. It is defined as per [I-D.ietf-bess-evpn-vpws-fxc].

6.2. Ethernet-tag Field vs AC ID Extended Community

The current proposal is entirely backward compabitible with [RFC7432] VLAN-aware bundling mode since the Ethernet-tag field remains intact. However, it has its own drawbacks. For instance with multicast, the same (S,G) maybe be used over different subnets. In that case, the same route MUST carry multiple AC ID Extended Community; one per attachment Circuit ID / VLAN. It may happen that the number of VLAN is faily large. Multiple routes with different RD may be required to carry such amount of Extended Community. This approach is complexifying the overall solution and implementation.

To remedy to that situation, the attachment Circuit ID MAY be set to 0xFFFF_FFFF. That value tells peer PE that the attachment Circuit ID is carried has part of the Ethernet Tag field of the associated route. Since the key of the EVPN route is unique, multiple AC ID Extended Community per route is no longer required. There is drawback. It pose backward interoperability issue with PE expecting a zero Ethernet-TAG ID.

7. Security Considerations

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

8. IANA Considerations

A new transitive extended community Type of 0x06 and Sub-Type of 0x0E for EVPN Attachment Circuit Extended Community has been allocated by IANA.

9. Acknowledgement

10. References

10.1. Normative 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>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

10.2. Informative References

[I-D.ietf-bess-evpn-vpws-fxc]
Sajassi, A., Brissette, P., Uttaro, J., Drake, J., Boutros, S., and J. Rabadan, "EVPN VPWS Flexible Cross-Connect Service", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-vpws-fxc-07, 16 June 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-vpws-fxc-07.txt>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[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>.

[RFC9136] Rabadan, J., Ed., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in Ethernet VPN (EVPN)", RFC 9136, DOI 10.17487/RFC9136, October 2021, <https://www.rfc-editor.org/info/rfc9136>.

[RFC9251] Sajassi, A., Thoria, S., Mishra, M., Patel, K., Drake, J., and W. Lin, "Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Proxies for Ethernet VPN (EVPN)", RFC 9251, DOI 10.17487/RFC9251, June 2022, <https://www.rfc-editor.org/info/rfc9251>.

Authors' Addresses

Ali Sajassi
Cisco Systems
Email: sajassi@cisco.com

Patrice Brissette
Cisco Systems

Email: pbrisset@cisco.com

Mankamana Mishra
Cisco Systems
Email: mankamis@cisco.com

Samir Thoria
Cisco Systems
Email: sthoria@cisco.com

Jorge Rabadan
Nokia
Email: jorge.rabadan@nokia.com

John Drake
Juniper Networks
Email: jdrake@juniper.net