BESS
Internet-Draft
Intended status: Standards Track
Expires: May 17, 2018

J. Heitz
A. Sajassi
Cisco
J. Drake
Juniper
J. Rabadan
Nokia

November 13, 2017

Multi-homing and E-Tree in EVPN with Inter-AS Option B
draft-heitz-bess-evpn-option-b-01

Abstract

The BGP speaker that originates an EVPN Ethernet A-D per ES route is identified by the next-hop of the route. When the route is propagated by an ASBR as an Inter-AS Option B route, the ASBR overwrites the next-hop. This document describes a method to identify the originator of the route.

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 [RFC2119].

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 May 17, 2018.

Table of Contents

1. Terminology

Inter-AS Option B: This is described in Section 10.b of [RFC4364]

EAD-per-ES: Ethernet A-D per Ethernet Segment Route.

EAD-per-EVI: Ethernet A-D per EVPN Instance Route.

EAD: EVPN Type 1 route: Ethernet Auto-discovery Route. Either an EAD-per-ES or an EAD-per-EVI route.

Type 2/5: either the EVPN Type 2 route: MAC/IP Advertisement Route or the EVPN Type 5 route: IP Prefix Route described in [I-D.ietf-bess-evpn-prefix-advertisement].

Mass Withdraw: To withdraw the route from the forwarding table. For example, a MAC route that is mass withdrawn remains in the BGP table. The MAC route is required for directing packets with the specified MAC destination address to a matching backup or alias route. When a MAC route is completely withdrawn, then the matching backup or alias routes can no longer be used for the given MAC address. The withdrawal of an EAD-per-ES route will cause the mass withdrawal of associated Type 2/5 routes as well as associated EAD-per-EVI routes.

2. Introduction

Inter-AS Option B is illustrated in Figure 1.

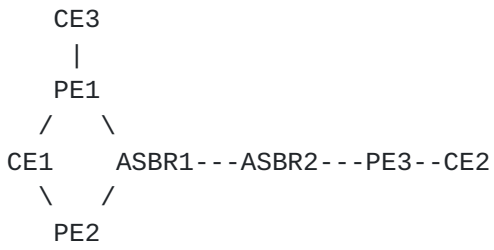

Figure 1: Inter-AS Option B

Traffic flow is from CE2 to CE1 where PE3 is an imposition PE, and PE1 and PE2 are disposition PEs. The following sections describe the issues that EVPN multi-homing and EVPN E-tree services have in these types of scenarios.

2.1. EVPN multi-homing and Inter-AS Option B issue

In a multi-homing scenario, the router that performs the redundancy switchover or the load balancing (e.g. PE3) must know which router originated the Ethernet A-D routes. These redundancy functions are normally implemented on a PE, but not on an ASBR.

Quote from [RFC7432]:

> "A remote PE that receives a MAC/IP Advertisement route with a non-reserved ESI SHOULD consider the advertised MAC address to be reachable via all PEs that have advertised reachability to that MAC address's EVI/ES via the combination of an Ethernet A-D per

EVI route for that EVI/ES (and Ethernet tag, if applicable) AND Ethernet A-D per ES routes for that ES."

In the Intra-AS case, the remote PE identifies the "PEs that have advertised reachability" by the next-hops of the Ethernet A-D routes. In the Inter-AS option B case, ASBR1 and ASBR2 rewrite the next-hops to themselves on all EVPN route advertisements, thus losing the identity of the PE that originated an advertisement.

As a result, PE3 is unable to distinguish an EAD-per-ES route that originated at PE1 from one that originated at PE2.

2.2. EVPN E-tree and Inter-AS Option B issue

As described in [EVPN-Etree], leaf-to-leaf BUM traffic filtering is always performed at the disposition PE and based on the Leaf Label. The Leaf Label can be downstream allocated (ingress replication) or upstream allocated (p2mp tunnels) and is advertised in an EAD-per-ES route with ESI-0. As in the multi-homing case, the PEs must identify the PE that originated a given EAD-per-ES route, for both cases, ingress replication or p2mp tunnels, so that the leaf-to-leaf BUM filtering can be successful.

If ingress-replication is used for BUM traffic, the ingress PE must identify the originator of the ESI-0 EAD-per-ES route, program the Leaf Label and push it on the stack when sending BUM Leaf traffic to the egress PE. However, this identification of the originating PE is not possible in Inter-AS option B scenarios where ASBRs rewrite the next-hops. For instance, assuming CE2 and CE3 (Figure 1) are connected to Leaf ACs, PE1 will advertise a Leaf Label in an EAD-per-ES route for ESI-0. When CE2 sends BUM traffic, PE3 will not know what Leaf Label to use for sending traffic to PE1.

Similarly, when PE3 uses non-segmented p2mp tunnels for BUM traffic, PE3 will upstream allocate a Leaf Label and advertise it in an EAD-per-ES route, so that when sending BUM traffic with a Leaf Label, PE1 can identify that is coming from a Leaf and not forward it to CE3.

In both cases, the current Intra-AS procedures do not allow to identify the originator of the EAD-per-ES routes and therefore egress BUM filtering for leaf-to-leaf is not possible when the Leaf ACs are located on different AS'es.

3. Solution using the Tunnel Encapsulation Attribute

The Tunnel Encapsulation Attribute is specified in [I-D.ietf-idr-tunnel-encaps]. A new TLV to identify the Originating PE is specified here. It is called OPE. The tunnel type for the OPE

(suggested value 15) is to be assigned by IANA. The OPE MUST contain the Remote Endpoint Sub-TLV. The OPE must be able to uniquely identify the PE of origin within all ASes that participate in an EVPN instance.

If a BGP speaker, such as a route reflector or an ASBR, is about to re-advertise a Type 2/5 or EAD route that does not have a OPE, and will change the next-hop of that route, then it MUST add one by putting the received next-hop into the Remote Endpoint Sub-TLV of the OPE. This will ensure that all originating EVPN routes carry the necessary information for imposition PEs to function properly for aliasing and mass withdraw.

Any router that re-advertises a route that contains a OPE may modify some TLVs in the Tunnel Encapsulation Attribute attribute. However, it MUST keep the OPE unchanged. Examples are ASBR1 and ASBR2 in Figure 1.

4. Operation

For an inter-AS option B scenario, when a PE receives EVPN route(s) with OPE from an ASBR, then everything works per [RFC7432] specification including both aliasing function and mass withdraw. i.e., the imposition PE (e.g., PE3) can process mass withdraw messages (Ethernet A-D per ES route). However, if a PE receives EVPN route(s) without a OPE from an ASBR, then the mass withdraw function operates in a degenerate mode where only Ethernet A-D per EVI route can be processed (for its corresponding MAC-VRF) but not Ethernet A-D per ES route (corresponding to all the impacted MAC-VRFs). The following sections detail the procedures associated with OPE processing.

5. Procedures at the Imposition PE

5.1. Primer for subsequent sections

When routes are being compared, they must exist in the same MAC-VRF and have the same non-reserved ESI. In addition, when Type 2/5 routes and EAD-per-EVI routes are being compared, they must have the same Ethernet Tag. Type 2/5 routes with ESI==0 do not use mass withdrawal or aliasing.

5.2. OPE exists on all Type 2/5 and EAD Routes

If all Type 2/5 and EAD routes have a OPE, then "PEs that have advertised reachability" can be identified by the OPE and the procedures of [RFC7432] can be applied without modification.

5.3. Some routes do not contain OPE

The routes that have a OPE are handled as per the previous section. The routes that do not have a OPE need the following procedures.

Type 2/5 routes without a OPE and EAD-per-EVI routes without a OPE are valid if at least one EAD-per-ES route without a OPE exists with the same next-hop. In other words: if multiple EAD-per-ES routes with the same next-hop as a Type 2/5 route exist, then the Type 2/5 route will only be mass withdrawn once all of the EAD-per-ES routes are withdrawn. This rule is necessary, because a BGP speaker may serve dual roles as ASBR and PE

[Editorial note: If it is determined that no BGP speakers exist that do not normally follow the procedures in this document (Legacy speakers) then the following sub sections may be omitted]

If an EAD-per-EVI route without a OPE is withdrawn, it will mass withdraw all Type 2/5 routes without a OPE that have the same next-hop and the same RD as the EAD-per-EVI route. This is called mass-withdraw per EVI. Note, it is not the absence of the EAD-per-EVI route that causes mass-withdrawal, but the actual withdrawal itself. If the route was never there to begin with, then no withdrawal took place.

If any entity in the network rewrites an RD, then all entities must rewrite the RD in a consistent manner, such that routes with the same RD continue to have the same RD and routes with different RDs continue to have different RDs. Note that if this condition is violated, then other network functions would also break.

5.4. OPE exists on EAD routes, but not on Type 2/5 routes

If a Type 2/5 route exists without a OPE and an EAD-per-EVI route exists with a OPE and it has the same next-hop and the same RD as the Type 2/5 route, then the Type 2/5 route shall inherit the OPE from the EAD-per-EVI route. Thereafter, Section 5.2 applies.

6. Security Considerations

TBD

7. IANA Considerations

A Tunnel Encapsulation Attribute Tunnel Type for the OPE is required.

8. Acknowledgements

Thanks to Kiran Pillai, Patrice Brissette, Satya Mohanty and Keyur Patel for careful review and suggestions.

9. Appendix

9.1. Alternative Ways to Signal OPE

[Note to RFC editor: This appendix to be removed before publication]

9.1.1. Extended Community holding the IP addres

The Extended Community to use must be transitive and either IPv4 Specific or IPv6 Specific as described in [RFC5701]. Thus, if it is IPv4 Specific, it will be of type 0x41 and if IPv6 Specific, it will be of type 0x40.

The Extended Community will hold the IP address of the PE that originates the EVPN routes.

9.1.2. Large Community holding the BGP Identifier

A PE can be uniquely identified by its BGP identifier (also called Router ID) and its AS number (ASN). A Large Community [RFC8092] can be used to carry the BGP identifier and the ASN. A well known Large Community needs to be allocated for this. This allocation is for the Global Administrator field. The Local Data Part 1 field should carry ASN and the Local Data Part 2 should carry the BGP identifier.

9.2. Considerations

It may be possible to associate the EAD-per-ES route with the Type 2/5 route by matching the Administrator Subfield of the RD. However, there are too many constraints that need to be met to make this method reliable. Basically, the RD was emphatically designed to distinguish routes, not to identify them. The constraints that need to be met are:

- The RD MUST by of Type 1. [RFC7432] recommends Type 1, but does not mandate it.

- The Administrator subfield of the RD MUST be the same for each of these routes originated by one PE. [RFC7432] does not require this. It just says "The value field comprises an IP address of the PE", but does not say that it must be the same IP address for all. In an IPv6 only scenario, other ways will be used to assign RD.

- The Administrator subfield of the RD MUST be unique among all PEs participating in the Inter-AS EVPN. This is likely, but not guaranteed.

- If RDs are rewritten at AS boundaries, then the Administrator subfield MUST be rewritten in a consistent way such as to preserve the above properties.

By allowing a single EAD-per-ES route to validate all EAD-per-EVI routes and all Type 2/5 routes, some of those routes may be falsely validated. However that is the best possible outcome without a OPE. It is transient until the Type 2/5 route can be withdrawn.

The possibility of the address space of PE next-hops in one AS overlapping that of another AS was raised. In such a case, the IP address of a PE in one AS may be the same as the IP address of a different PE in another AS. Because an ASBR overwrites next-hops, this can work. The OPE contains both the ASN as well as the IP address of the originating PE, so this works too. However, EVPN route types 3 and 4 contain only the originating router's IP address, but not the originating router's ASN. Therefore, EVPN route types 3 and 4 may also need a OPE.

The possibility of making the EAD-per-EVI route mandatory was raised. This would make some of the procedures easier, because the RD of the EAD-per-EVI route can be matched with the RD of the Type 2/5 route

10. Normative References

[I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Palislamovic, S., and A. Isaac, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-02 (work in progress), September 2015.

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-02 (work in progress), May 2016.

[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>.

[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>.

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, DOI 10.17487/RFC5701, November 2009, <https://www.rfc-editor.org/info/rfc5701>.

[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>.

[RFC8092] Heitz, J., Ed., Snijders, J., Ed., Patel, K., Bagdonas, I., and N. Hilliard, "BGP Large Communities Attribute", RFC 8092, DOI 10.17487/RFC8092, February 2017, <https://www.rfc-editor.org/info/rfc8092>.

Authors' Addresses

Jakob Heitz
Cisco
170 West Tasman Drive
San Jose, CA 95134
USA

Email: jheitz@cisco.com

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134
USA

Email: sajassi@cisco.com

John Drake
Juniper

Email: jdrake@juniper.net

Jorge Rabadan
Nokia

Email: jorge.rabadan@nokia.com