BESS
Internet-Draft
Updates: 7432 (if approved)
Intended status: Standards Track
Expires: May 11, 2022

Z. Zhang
W. Lin
Juniper Networks
J. Rabadan
Nokia
K. Patel
Arrcus
A. Sajassi
Cisco Systems

November 7, 2021

Updates on EVPN BUM Procedures
draft-ietf-bess-evpn-bum-procedure-updates-13

Abstract

This document specifies updated procedures for handling broadcast, unknown unicast, and multicast (BUM) traffic in Ethernet VPNs (EVPN), including selective multicast, and provider tunnel segmentation. This document updates RFC 7432.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 11, 2022.

Table of Contents

1. Terminology 3
2. Introduction 3
2.1. Tunnel Segmentation 4
2.1.1. Reasons for Tunnel Segmentation 5
3. Additional Route Types of EVPN NLRI 6
3.1. Per-Region I-PMSI A-D route 7
3.2. S-PMSI A-D route 7
3.3. Leaf A-D route 8
4. Selective Multicast 8
5. Inter-AS Segmentation 9
5.1. Differences from Section 7.2.2 of [RFC7117] When Applied to EVPN 9
5.2. I-PMSI Leaf Tracking 11
5.3. Backward Compatibility 11
5.3.1. Designated ASBR Election 13
6. Inter-Region Segmentation 13
6.1. Area/AS vs. Region 13
6.2. Per-region Aggregation 14
6.3. Use of S-NH-EC 15
6.4. Ingress PE's I-PMSI Leaf Tracking 16
7. Multi-homing Support 16
8. IANA Considerations 17
9. Security Considerations 17
10. Acknowledgements 17
11. Contributors 17
12. References 18
12.1. Normative References 18
12.2. Informative References 19
Authors' Addresses 20

1. Terminology

It is expected that audience is familiar with MVPN [RFC6513] [RFC6514], VPLS Multicast [RFC7117] and EVPN [RFC7432] concepts and terminologies. For convenience, the following terms are briefly explained.

- PMSI [RFC6513]: P-Multicast Service Interface - a conceptual interface for a PE to send customer multicast traffic to all or some PEs in the same VPN.
- I-PMSI: Inclusive PMSI - to all PEs in the same VPN.
- S-PMSI: Selective PMSI - to some of the PEs in the same VPN.
- I/S-PMSI A-D Route: Auto-Discovery routes used to announce the tunnels that instantiate an I/S-PMSI.
- Leaf Auto-Discovery (A-D) routes [RFC6513]: For explicit leaf tracking purpose. Triggered by I/S-PMSI A-D routes and targeted at triggering route's (re-)advertiser. Its NLRI embeds the entire NLRI of the triggering PMSI A-D route.
- IMET A-D route [RFC7432]: Inclusive Multicast Ethernet Tag A-D route. The EVPN equivalent of MVPN Intra-AS I-PMSI A-D route used to announce the tunnels that instantiate an I-PMSI.
- SMET A-D route [I-D.ietf-bess-evpn-igmp-mld-proxy]: Selective Multicast Ethernet Tag A-D route. The EVPN equivalent of MVPN Leaf A-D route but unsolicited and untargeted.
- PMSI Tunnel Attribute (PTA): An optional transitive BGP attribute that may be attached to PMSI/Leaf A-D routes to provide information for a PMSI tunnel.

2. Introduction

[RFC7117] specifies procedures for Multicast in Virtual Private LAN Service (VPLS Multicast) using both inclusive tunnels and selective tunnels with or without inter-as segmentation, similar to the Multicast VPN (MVPN) procedures specified in [RFC6513] and [RFC6514]. [RFC7524] specifies inter-area tunnel segmentation procedures for both VPLS Multicast and MVPN.

[RFC7432] specifies BGP MPLS-Based Ethernet VPN (EVPN) procedures, including those handling broadcast, unknown unicast, and multicast (BUM) traffic. A lot of details are referred to [RFC7117], yet with

quite some feature gaps like selective tunnel and tunnel segmentation (Section 2.1).

This document aims at filling the gaps - cover the use of selective and segmented tunnels in EVPN. It follows the same editorial choice as in RFC7432 and only specifies differences from relevant procedures in [RFC7117] and [RFC7524], instead of repeating the text. Note that these differences are applicable to EVPN only, and are not updates to [RFC7117] or [RFC7524].

MVPN, VPLS and EVPN all have the need to discover other PEs in the same L3/L2 VPN and announce the inclusive tunnels. MVPN introduced the I-PMSI concept and uses I-PMSI A-D route for that. EVPN uses Inclusive Multicast Ethernet Tag Route (IMET) A-D route but VPLS just adds an PMSI Tunnel Attribute (PTA) to the existing VPLS A-D route for that purpose. For selective tunnels, they all do use the same term S-PMSI A-D routes.

Many places of this document involve the I-PMSI concept that is all the same for all three technologies. For consistency and convenience, EVPN's IMET and VPLS's VPLS A-D route carrying PTA for BUM traffic purpose may all be referred to as I-PMSI A-D routes depending on the context.

2.1. Tunnel Segmentation

MVPN provider tunnels and EVPN/VPLS BUM provider tunnels, which are referred to as MVPN/EVPN/VPLS provider tunnels in this document for simplicity, can be segmented for technical or administrative reasons, which are summarized in Section 2.1.1 of this document. [RFC6513] and [RFC6514] cover MVPN inter-as segmentation, [RFC7117] covers VPLS multicast inter-as segmentation, and [RFC7524] (Seamless MPLS Multicast) covers inter-area segmentation for both MVPN and VPLS.

With tunnel segmentation, different segments of an end-to-end tunnel may have different encapsulation overhead. However, the largest overhead of the tunnel caused by an encapsulation method on a particular segment is not different from the case of a non-segmented tunnel with that encapsulation method. This is similar to the case of a network with different link types.

There is a difference between MVPN and VPLS multicast inter-as segmentation (the VPLS approach is briefly discribed in Section 5.1). For simplicity, EVPN will use the same procedures as in MVPN. All ASBRs can re-advertise their choice of the best route. Each can become the root of its intra-AS segment and inject traffic it receives from its upstream, while each downstream PE/ASBR will only

pick one of the upstream ASBRs as its upstream. This is also the behavior even for VPLS in case of inter-area segmentation.

For inter-area segmentation, [RFC7524] requires the use of Inter-area P2MP Segmented Next-Hop Extended Community (S-NH-EC), and the setting of "Leaf Information Required" L flag in PTA in certain situations. In the EVPN case, the requirements around S-NH-EC and the PTA "L" flag differ from [RFC7524] to make the segmentation procedures transparent to ingress and egress PEs.

[RFC7524] assumes that segmentation happens at area borders. However, it could be at "regional" borders, where a region could be a sub-area, or even an entire AS plus its external links (Section 6.1). That would allow for more flexible deployment scenarios (e.g. for single-area provider networks). This document extends the inter-area segmentation to inter-region segmentation for EVPN.

2.1.1. Reasons for Tunnel Segmentation

Tunnel segmentation may be required and/or desired because of administrative and/or technical reasons.

For example, an MVPN/VPLS/EVPN network may span multiple providers and the end-to-end provider tunnels have to be segmented at and stitched by the ASBRs. Different providers may use different tunnel technologies (e.g., provider A uses Ingress Replication [RFC7988], provider B uses RSVP-TE P2MP [RFC4875] while provider C uses mLDP [RFC6388]). Even if they use the same tunnel technology like RSVP-TE P2MP, it may be impractical to set up the tunnels across provider boundaries.

The same situations may apply between the ASes and/or areas of a single provider. For example, the backbone area may use RSVP-TE P2MP tunnels while non-backbone areas may use mLDP tunnels.

Segmentation can also be used to divide an AS/area into smaller regions, so that control plane state and/or forwarding plane state/ burden can be limited to that of individual regions. For example, instead of Ingress Replicating to 100 PEs in the entire AS, with inter-area segmentation [RFC7524] a PE only needs to replicate to local PEs and ABRs. The ABRs will further replicate to their downstream PEs and ABRs. This not only reduces the forwarding plane burden, but also reduces the leaf tracking burden in the control plane.

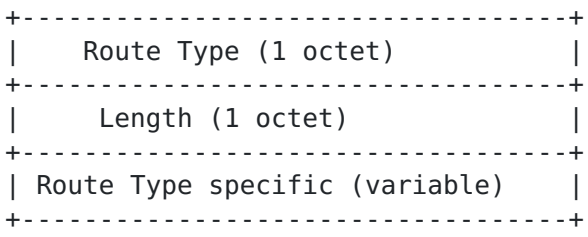
Smaller regions also have the benefit that, in case of tunnel aggregation, it is easier to find congruence among the segments of different constituent (service) tunnels and the resulting aggregation

(base) tunnel in a region. This leads to better bandwidth efficiency, because the more congruent they are, the fewer leaves of the base tunnel need to discard traffic when a service tunnel's segment does not need to receive the traffic (yet it is receiving the traffic due to aggregation).

Another advantage of the smaller region is smaller BIER [RFC8279] sub-domains. With BIER, packets carry a BitString, in which the bits correspond to edge routers that needs to receive traffic. Smaller sub-domains means smaller BitStrings can be used without having to send multiple copies of the same packet.

3. Additional Route Types of EVPN NLRI

[RFC7432] defines the format of EVPN NLRI as the following:



So far eight route types have been defined in [RFC7432], [I-D.ietf-bess-evpn-prefix-advertisement], and [I-D.ietf-bess-evpn-igmp-mld-proxy]:

+ 1 - Ethernet Auto-Discovery (A-D) route
+ 2 - MAC/IP Advertisement route
+ 3 - Inclusive Multicast Ethernet Tag route
+ 4 - Ethernet Segment route
+ 5 - IP Prefix Route
+ 6 - Selective Multicast Ethernet Tag Route
+ 7 - Multicast Join Synch Route
+ 8 - Multicast Leave Synch Route

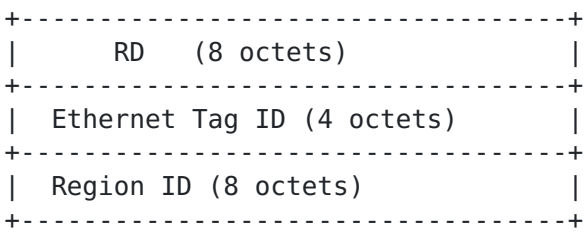
This document defines three additional route types:

+ 9 - Per-Region I-PMSI A-D route
+ 10 - S-PMSI A-D route
+ 11 - Leaf A-D route

The "Route Type specific" field of the type 9 and type 10 EVPN NLRIs starts with a type 1 RD, whose Administrator sub-field MUST match that of the RD in all current non-Leaf A-D (Section 3.3) EVPN routes from the same advertising router for a given EVI.

3.1. Per-Region I-PMSI A-D route

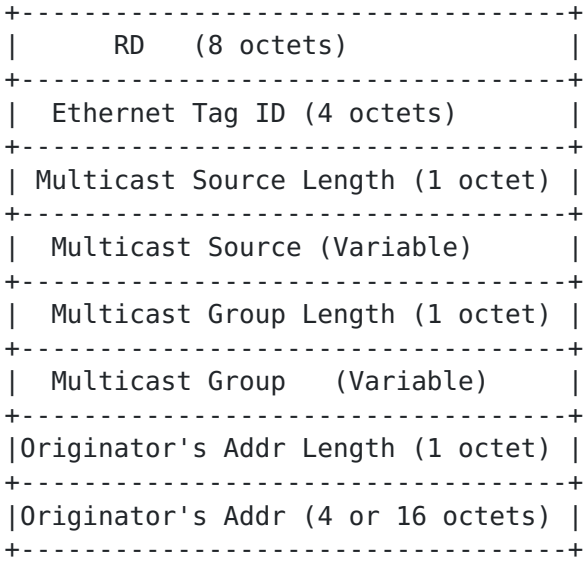
The Per-region I-PMSI A-D route has the following format. Its usage is discussed in Section 6.2.



The Region ID identifies the region and is encoded just as how an Extended Community is encoded, as detailed in Section 6.2.

3.2. S-PMSI A-D route

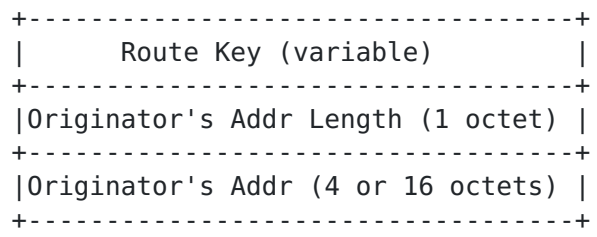
The S-PMSI A-D route has the following format:



Other than the addition of Ethernet Tag ID and Originator's Addr Length, it is identical to the S-PMSI A-D route as defined in [RFC7117]. The procedures in [RFC7117] also apply (including wildcard functionality), except that the granularity level is per Ethernet Tag.

3.3. Leaf A-D route

The Route Type specific field of a Leaf A-D route consists of the following:



A Leaf A-D route is originated in response to a PMSI route, which could be an Inclusive Multicast Tag route, a per-region I-PMSI A-D route, an S-PMSI A-D route, or some other types of routes that may be defined in the future that triggers Leaf A-D routes. The Route Key is the NLRI of the route for which this Leaf A-D route is generated.

The general procedures of Leaf A-D route are first specified in [RFC6514] for MVPN. The principles apply to VPLS and EVPN as well. [RFC7117] has details for VPLS Multicast, and this document points out some specifics for EVPN, e.g. in Section 5.

4. Selective Multicast

[I-D.ietf-bess-evpn-igmp-mld-proxy] specifies procedures for EVPN selective forwarding of IP multicast using SMET routes. It assumes selective forwarding is always used with IR for all flows (though the same signaling can also be used for an ingress PE to find out the set of egress PEs for selective forwarding with BIER). An NVE proxies the IGMP/MLD state that it learns on its ACs to (C-S,C-G) or (C-*,C-G) SMET routes that advertises to other NVEs, and a receiving NVE converts the SMET routes back to IGMP/MLD messages and sends them out of its ACs. The receiving NVE also uses the SMET routes to identify which NVEs need to receive traffic for a particular (C-S,C-G) or (C-*,C-G) to achieve selective forwarding using IR or BIER.

With the above procedures, selective forwarding is done for all flows and the SMET routes are advertised for all flows. It is possible that an operator may not want to track all those (C-S, C-G) or (C-*,C-G) state on the NVEs, and the multicast traffic pattern allows inclusive forwarding for most flows while selective forwarding is needed only for a few high-rate flows. For that, or for tunnel types other than IR/BIER, S-PMSI/Leaf A-D procedures defined for Selective Multicast for VPLS in [RFC7117] are used. Other than that different route types and formats are specified with EVPN SAFI for S-PMSI A-D

and Leaf A-D routes (Section 3), all procedures in [RFC7117] with respect to Selective Multicast apply to EVPN as well, including wildcard procedures. In a nutshell, a source NVE advertises S-PMSI A-D routes to announce the tunnels used for certain flows, and receiving NVEs either join the announced PIM/mLDP tunnel or respond with Leaf A-D routes if the Leaf Information Required flag is set in the S-PMSI A-D route's PTA (so that the source NVE can include them as tunnel leaves).

An optimization to the [RFC7117] procedures may be applied. Even if a source NVE sets the L flag to request Leaf A-D routes, an egress NVE MAY omit the Leaf A-D route if it has already advertised a corresponding SMET route, and the source NVE MUST use that in lieu of the Leaf A-D route.

The optional optimizations specified for MVPN in [RFC8534] are also applicable to EVPN when the S-PMSI/Leaf A-D routes procedures are used for EVPN selective multicast forwarding.

5. Inter-AS Segmentation

5.1. Differences from Section 7.2.2 of [RFC7117] When Applied to EVPN

The first paragraph of Section 7.2.2.2 of [RFC7117] says:

"... The best route procedures ensure that if multiple ASBRs, in an AS, receive the same Inter-AS A-D route from their EBGP neighbors, only one of these ASBRs propagates this route in Internal BGP (IBGP). This ASBR becomes the root of the intra-AS segment of the inter-AS tree and ensures that this is the only ASBR that accepts traffic into this AS from the inter-AS tree."

The above VPLS behavior requires complicated VPLS specific procedures for the ASBRs to reach agreement. For EVPN, a different approach is used and the above quoted text is not applicable to EVPN.

With the different approach for EVPN/MVPN, each ASBR will re-advertise its received Inter-AS A-D route to its IBGP peers and becomes the root of an intra-AS segment of the inter-AS tree. The intra-AS segment rooted at one ASBR is disjoint with another intra-AS segment rooted at another ASBR. This is the same as the procedures for S-PMSI in [RFC7117] itself.

The following bullet in Section 7.2.2.2 of [RFC7117] does not apply to EVPN.

+ If the ASBR uses ingress replication to instantiate the intra-AS segment of the inter-AS tunnel, the re-advertised route MUST NOT carry the PMSI Tunnel attribute.

The following bullet in Section 7.2.2.2 of [RFC7117]:

+ If the ASBR uses a P-multicast tree to instantiate the intra-AS segment of the inter-AS tunnel, the PMSI Tunnel attribute MUST contain the identity of the tree that is used to instantiate the segment (note that the ASBR could create the identity of the tree prior to the actual instantiation of the segment). If, in order to instantiate the segment, the ASBR needs to know the leaves of the tree, then the ASBR obtains this information from the A-D routes received from other PEs/ASBRs in the ASBR's own AS.

is changed to the following when applied to EVPN:

"The PMSI Tunnel attribute MUST specify the tunnel for the segment. If and only if, in order to establish the tunnel, the ASBR needs to know the leaves of the tree, then the ASBR MUST set the L flag to 1 in the PTA to trigger Leaf A-D routes from egress PEs and downstream ASBRs. It MUST be (auto-)configured with an import RT, which controls acceptance of leaf A-D routes by the ASBR."

Accordingly, the following paragraph in Section 7.2.2.4 of [RFC7117]:

"If the received Inter-AS A-D route carries the PMSI Tunnel attribute with the Tunnel Identifier set to RSVP-TE P2MP LSP, then the ASBR that originated the route MUST establish an RSVP-TE P2MP LSP with the local PE/ASBR as a leaf. This LSP MAY have been established before the local PE/ASBR receives the route, or it MAY be established after the local PE receives the route."

is changed to the following when applied to EVPN:

"If the received Inter-AS A-D route has the L flag set in its PTA, then a receiving PE MUST originate a corresponding Leaf A-D route, while a receiving ASBR MUST originate a corresponding Leaf A-D route if and only if it received and imported one or more corresponding Leaf A-D routes from its downstream IBGP or EBGP peers, or it has non-null downstream forwarding state for the PIM/mLDP tunnel that instantiates its downstream intra-AS segment. The targeted ASBR for the Leaf A-D route, which (re-)advertised the Inter-AS A-D route, MUST establish a tunnel to the leaves discovered by the Leaf A-D routes."

5.2. I-PMSI Leaf Tracking

An ingress PE does not set the L flag in its Inclusive Multicast Ethernet Tag (IMET) A-D route's PTA, even with Ingress Replication or RSVP-TE P2MP tunnels. It does not rely on the Leaf A-D routes to discover leaves in its AS, and Section 11.2 of [RFC7432] explicitly states that the L flag must be set to zero.

An implementation of [RFC7432] might have used the Originating Router's IP Address field of the IMET A-D routes to determine the leaves, or might have used the Next Hop field instead. Within the same AS, both will lead to the same result.

With segmentation, an ingress PE MUST determine the leaves in its AS from the BGP next hops in all its received IMET A-D routes, so it does not have to set the L flag set to request Leaf A-D routes. PEs within the same AS will all have different next hops in their IMET A-D routes (hence will all be considered as leaves), and PEs from other ASes will have the next hop in their IMET A-D routes set to addresses of ASBRs in this local AS, hence only those ASBRs will be considered as leaves (as proxies for those PEs in other ASes). Note that in case of Ingress Replication, when an ASBR re-advertises IMET A-D routes to IBGP peers, it MUST advertise the same label for all those for the same Ethernet Tag ID and the same EVI. Otherwise, duplicated copies will be sent by the ingress PE and received by egress PEs in other regions. For the same reason, when an ingress PE builds its flooding list, if multiple routes have the same (nexthop, label) tuple they MUST only be added as a single branch in the flooding list.

5.3. Backward Compatibility

The above procedures assume that all PEs are upgraded to support the segmentation procedures:

- An ingress PE uses the Next Hop and not Originating Router's IP Address to determine leaves for the I-PMSI tunnel.

- An egress PE sends Leaf A-D routes in response to I-PMSI routes, if the PTA has the L flag set by the re-advertising ASBR.

- In case of Ingress Replication, when an ingress PE builds its flooding list, multiple I-PMSI routes may have the same (nexthop, label) tuple and only a single branch for those will be added in the flooding list.

If a deployment has legacy PEs that does not support the above, then a legacy ingress PE would include all PEs (including those in remote

ASes) as leaves of the inclusive tunnel and try to send traffic to them directly (no segmentation), which is either undesired or not possible; a legacy egress PE would not send Leaf A-D routes so the ASBRs would not know to send external traffic to them.

If this backward compatibility problem needs to be addressed, the following procedure MUST be used (see Section 6.2 for per-PE/AS/region I-PMSI A-D routes):

- An upgraded PE indicates in its per-PE I-PMSI A-D route that it supports the new procedures. This is done by setting a flag bit in the EVPN Multicast Flags Extended Community.

- All per-PE I-PMSI A-D routes are restricted to the local AS and not propagated to external peers.

- The ASBRs in an AS originate per-region I-PMSI A-D routes and advertise them to their external peers to specify tunnels used to carry traffic from the local AS to other ASes. Depending on the types of tunnels being used, the L flag in the PTA may be set, in which case the downstream ASBRs and upgraded PEs will send Leaf A-D routes to pull traffic from their upstream ASBRs. In a particular downstream AS, one of the ASBRs is elected, based on the per-region I-PMSI A-D routes for a particular source AS, to send traffic from that source AS to legacy PEs in the downstream AS. The traffic arrives at the elected ASBR on the tunnel announced in the best per-region I-PMSI A-D route for the source AS, that the ASBR has selected of all those that it received over EBGP or IBGP sessions. The election procedure is described in Section 5.3.1.

- In an ingress/upstream AS, if and only if an ASBR has active downstream receivers (PEs and ASBRs), which are learned either explicitly via Leaf A-D routes or implicitly via PIM join or mLDP label mapping, the ASBR originates a per-PE I-PMSI A-D route (i.e., regular Inclusive Multicast Ethernet Tag route) into the local AS, and stitches incoming per-PE I-PMSI tunnels into its per-region I-PMSI tunnel. With this, it gets traffic from local PEs and send to other ASes via the tunnel announced in its per-region I-PMSI A-D route.

Note that, even if there is no backward compatibility issue, the use of per-region I-PMSI has the benefit of keeping all per-PE I-PMSI A-D routes in their local ASes, greatly reducing the flooding of the routes and their corresponding Leaf A-D routes (when needed), and the number of inter-as tunnels.

5.3.1. Designated ASBR Election

When an ASBR re-advertises a per-region I-PMSI A-D route into an AS in which a designated ASBR needs to be used to forward traffic to the legacy PEs in the AS, it MUST include a DF Election EC. The EC and its use is specified in [RFC8584]. The AC-DF bit in the DF Election EC MUST be cleared. If it is known that no legacy PEs exist in the AS, the ASBR MUST NOT include the EC and MUST remove the DF Election EC if one is carried in the per-region I-PMSI A-D routes that it receives. Note that this is done for each set of per-region I-PMSI A-D routes with the same NLRI.

Based on the procedures in [RFC8584], an election algorithm is determined according to the DF Election ECs carried in the set of per-region I-PMSI routes of the same NLRI re-adverised into the AS. The algorithm is then applied to a candidate list, which is the set of ASBRs that re-advertised the per-region I-PMSI routes of the same NLRI carrying the DF Election EC.

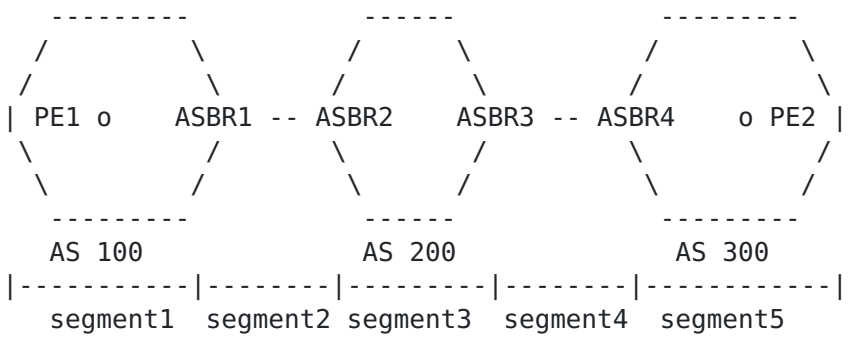
6. Inter-Region Segmentation

6.1. Area/AS vs. Region

[RFC7524] is for MVPN/VPLS inter-area segmentation and does not explicitly cover EVPN. However, if "area" is replaced by "region" and "ABR" is replaced by "RBR" (Regional Border Router) then everything still works, and can be applied to EVPN as well.

A region can be a sub-area, or can be an entire AS including its external links. Instead of automatic region definition based on IGP areas, a region would be defined as a BGP peer group. In fact, even with IGP area based region definition, a BGP peer group listing the PEs and ABRs in an area is still needed.

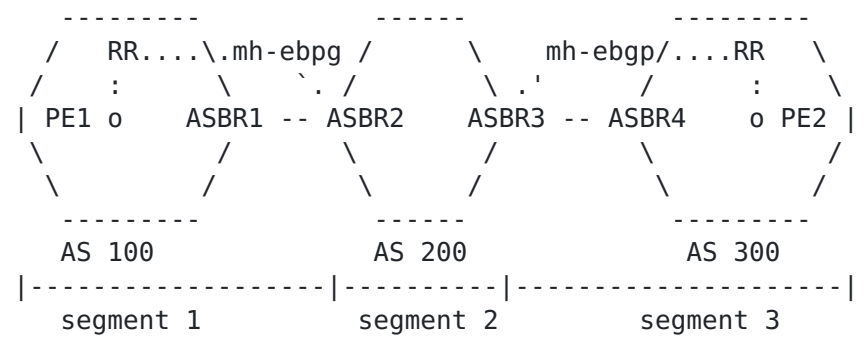
Consider the following example diagram for inter-as segmentation:



The inter-as segmentation procedures specified so far ([RFC6513] [RFC6514], [RFC7117], and Section 5 of this document) require all ASBRs to be involved, and Ingress Replication is used between two ASBRs in different ASes.

In the above diagram, it's possible that ASBR1/4 does not support segmentation, and the provider tunnels in AS 100/300 can actually extend across the external link. In this case, the inter-region segmentation procedures can be used instead - a region is the entire (AS100 + ASBR1-ASBR2 link) or (AS300 + ASBR3-ASBR4 link). ASBR2/3 would be the RBRs, and ASBR1/4 will just be a transit core router with respect to provider tunnels.

As illustrated in the diagram below, ASBR2/3 will establish a multihop EBGP session with either a RR or directly with PEs in the neighboring AS. I/S-PMSI A-D routes from ingress PEs will not be processed by ASBR1/4. When ASBR2 re-advertises the routes into AS 200, it changes the next hop to its own address and changes PTA to specify the tunnel type/identification in its own AS. When ASBR3 re-advertises I/S-PMSI A-D routes into the neighboring AS 300, it changes the next hop to its own address and changes PTA to specify the tunnel type/identification in the neighboring region. Now the segment is rooted at ASBR3 and extends across the external link to PEs.



6.2. Per-region Aggregation

Notice that every I/S-PMSI route from each PE will be propagated throughout all the ASes or regions. They may also trigger corresponding Leaf A-D routes depending on the types of tunnels used in each region. This may become too many - routes and corresponding tunnels. To address this concern, the I-PMSI routes from all PEs in a AS/region can be aggregated into a single I-PMSI route originated from the RBRs, and traffic from all those individual I-PMSI tunnels will be switched into the single I-PMSI tunnel. This is like the MVPN Inter-AS I-PMSI route originated by ASBRs.

The MVPN Inter-AS I-PMSI A-D route can be better called as per-AS I-PMSI A-D route, to be compared against the (per-PE) Intra-AS I-PMSI A-D routes originated by each PE. In this document we will call it as per-region I-PMSI A-D route, in case we want to apply the aggregation at regional level. The per-PE I-PMSI routes will not be propagated to other regions. If multiple RBRs are connected to a region, then each will advertise such a route, with the same Region ID and Ethernet Tag ID (Section 3.1). Similar to the per-PE I-PMSI A-D routes, RBRs/PEs in a downstream region will each select a best one from all those re-advertised by the upstream RBRs, hence will only receive traffic injected by one of them.

MVPN does not aggregate S-PMSI routes from all PEs in an AS like it does for I-PMSIs routes, because the number of PEs that will advertise S-PMSI routes for the same (s,g) or (*,g) is small. This is also the case for EVPN, i.e., there is no per-region S-PMSI routes.

Notice that per-region I-PMSI routes can also be used to address backwards compatibility issue, as discussed in Section 5.3.

The Region ID in the per-region I-PMSI route's NLRI is encoded like an EC. For example, the Region ID can encode an AS number or area ID in the following EC format:

- For a two-octet AS number, a Transitive Two-Octet AS-Specific EC of sub-type 0x09 (Source AS), with the Global Administrator sub-field set to the AS number and the Local Administrator sub-field set to 0.

- For a four-octet AS number, a Transitive Four-Octet AS-Specific EC of sub-type 0x09 (Source AS), with the Global Administrator sub-field set to the AS number and the Local Administrator sub-field set to 0.

- For an area ID, a Transitive IPv4-Address-Specific EC of any sub-type, with the Global Administrator sub-field set to the area ID and the Local Administrator sub-field set to 0.

Uses of other EC encoding MAY be allowed as long as it uniquely identifies the region and the RBRs for the same region uses the same Region ID.

6.3. Use of S-NH-EC

[RFC7524] specifies the use of S-NH-EC because it does not allow ABRs to change the BGP next hop when they re-advertise I/S-PMSI A-D routes to downstream areas. That is only to be consistent with the MVPN

Inter-AS I-PMSI A-D routes, whose next hop must not be changed when they're re-advertised by the segmenting ABRs for reasons specific to MVPN. For EVPN, it is perfectly fine to change the next hop when RBRs re-advertise the I/S-PMSI A-D routes, instead of relying on S-NH-EC. As a result, this document specifies that RBRs change the BGP next hop when they re-advertise I/S-PMSI A-D routes and do not use S-NH-EC. The advantage of this is that neither ingress nor egress PEs need to understand/use S-NH-EC, and a consistent procedure (based on BGP next hop) is used for both inter-as and inter-region segmentation.

If a downstream PE/RBR needs to originate Leaf A-D routes, it constructs an IP-based Route Target Extended Community by placing the IP address carried in the Next Hop of the received I/S-PMSI A-D route in the Global Administrator field of the Community, with the Local Administrator field of this Community set to 0 and setting the Extended Communities attribute of the Leaf A-D route to that Community.

Similar to [RFC7524], the upstream RBR MUST (auto-)configure a RT with the Global Administrator field set to the Next Hop in the re-advertised I/S-PMSI A-D route and with the Local Administrator field set to 0. With this, the mechanisms specified in [RFC4684] for constrained BGP route distribution can be used along with this specification to ensure that only the needed PE/ABR will have to process a said Leaf A-D route.

6.4. Ingress PE's I-PMSI Leaf Tracking

[RFC7524] specifies that when an ingress PE/ASBR (re-)advertises an VPLS I-PMSI A-D route, it sets the L flag to 1 in the route's PTA. Similar to the inter-as case, this is actually not really needed for EVPN. To be consistent with the inter-as case, the ingress PE does not set the L flag in its originated I-PMSI A-D routes, and determines the leaves based on the BGP next hops in its received I-PMSI A-D routes, as specified in Section 5.2.

The same backward compatibility issue exists, and the same solution as in the inter-as case applies, as specified in Section 5.3.

7. Multi-homing Support

To support multi-homing with segmentation, ESI labels SHOULD be allocated from "Domain-wide Common Block" (DCB) [I-D.ietf-bess-mvpn-evpn-aggregation-label] for all tunnel types including Ingress Replication. Via means outside the scope of this document, PEs know that ESI labels are from DCB and then existing

multi-homing procedures work as is (whether a multi-homed Ethernet Segment spans across segmentation regions or not).

Not using DCB-allocated ESI labels is outside the scope of this document.

8. IANA Considerations

IANA has temporarily assigned the following new EVPN route types:

- 9 - Per-Region I-PMSI A-D route
- 10 - S-PMSI A-D route
- 11 - Leaf A-D route

This document requests IANA to assign one flag bit from the EVPN Multicast Flags Extended Community to be created in [I-D.draft-ietf-bess-evpn-igmp-mld-proxy]:

- Bit-S - The router supports segmentation procedure defined in this document

9. Security Considerations

The Selective Forwarding procedures via S-PMSI/Leaf A-D routes in this document are based on the same procedures for MVPN [RFC6513] [RFC6514] and VPLS Multicast [RFC7117]. The tunnel segmentation procedures in this document are based on the similar procedures for MVPN inter-AS [RFC6514] and inter-area [RFC7524] tunnel segmentation, and procedures for VPLS Multicast [RFC7117] inter-as tunnel segmentation. When applied to EVPN, they do not introduce new security concerns besides what have been discussed in [RFC6513], [RFC6514], [RFC7117], and [RFC7524]. They also do not introduce new security concerns compared to [RFC7432].

10. Acknowledgements

The authors thank Eric Rosen, John Drake, and Ron Bonica for their comments and suggestions.

11. Contributors

The following also contributed to this document through their earlier work in EVPN selective multicast.

Junlin Zhang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jackey.zhang@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

12. References

12.1. Normative References

[I-D.ietf-bess-evpn-igmp-mld-proxy]
Sajassi, A., Thoria, S., Mishra, M., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-14 (work in progress), October 2021.

[I-D.ietf-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", draft-ietf-bess-mvpn-evpn-aggregation-label-06 (work in progress), April 2021.

[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>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC7117] Aggarwal, R., Ed., Kamite, Y., Fang, L., Rekhter, Y., and C. Kodeboniya, "Multicast in Virtual Private LAN Service (VPLS)", RFC 7117, DOI 10.17487/RFC7117, February 2014, <https://www.rfc-editor.org/info/rfc7117>.

[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>.

[RFC7524] Rekhter, Y., Rosen, E., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area Point-to-Multipoint (P2MP) Segmented Label Switched Paths (LSPs)", RFC 7524, DOI 10.17487/RFC7524, May 2015, <https://www.rfc-editor.org/info/rfc7524>.

[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>.

[RFC8534] Dolganow, A., Kotalwar, J., Rosen, E., Ed., and Z. Zhang, "Explicit Tracking with Wildcard Routes in Multicast VPN", RFC 8534, DOI 10.17487/RFC8534, February 2019, <https://www.rfc-editor.org/info/rfc8534>.

[RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <https://www.rfc-editor.org/info/rfc8584>.

12.2. Informative References

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

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, DOI 10.17487/RFC4875, May 2007, <https://www.rfc-editor.org/info/rfc4875>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[RFC7988] Rosen, E., Ed., Subramanian, K., and Z. Zhang, "Ingress Replication Tunnels in Multicast VPN", RFC 7988, DOI 10.17487/RFC7988, October 2016, <https://www.rfc-editor.org/info/rfc7988>.

[RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <https://www.rfc-editor.org/info/rfc8279>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Wen Lin
Juniper Networks

EMail: wlin@juniper.net

Jorge Rabadan
Nokia

EMail: jorge.rabadan@nokia.com

Keyur Patel
Arrcus

EMail: keyur@arrcus.com

Ali Sajassi
Cisco Systems

EMail: sajassi@cisco.com