Updates on EVPN BUM Procedures
draft-ietf-bess-evpn-bum-procedure-updates-04

Abstract

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

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

Copyright Notice

Table of Contents

1. **Terminology**

    To be added

2. **Introduction**

    RFC 7432 specifies procedures to handle broadcast, unknown unicast, and multicast (BUM) traffic in Section 11, 12 and 16, using Inclusive Multicast Ethernet Tag Route. A lot of details are referred to RFC 7117 (VPLS Multicast). In particular, selective multicast is briefly mentioned for Ingress Replication but referred to RFC 7117.

    RFC 7117 specifies procedures for using both inclusive tunnels and selective tunnels, similar to MVPN procedures specified in RFC 6513 and RFC 6514. A new SAFI "MCAST-VPLS" is introduced, with two types of NLRIs that match MVPN's S-PMSI A-D routes and Leaf A-D routes. The same procedures can be applied to EVPN selective multicast for both Ingress Replication and other tunnel types, but new route types need to be defined under the same EVPN SAFI.

    MVPN uses terms I-PMSI and S-PMSI A-D Routes. For consistency and convenience, this document will use the same I/S-PMSI terms for VPLS and EVPN. In particular, EVPN's Inclusive Multicast Ethernet Tag Route and VPLS's VPLS A-D route carrying PTA (PMSI Tunnel Attribute) for BUM traffic purpose will all be referred to as I-PMSI A-D routes. Depending on the context, they may be used interchangeably.

    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 of this document. RFC 6513/6514 cover MVPN inter-as segmentation, RFC 7117 covers VPLS multicast inter-as segmentation, and RFC 7524 (Seamless MPLS Multicast) covers inter-area segmentation for both MVPN and VPLS.

    There is a difference between MVPN and VPLS multicast inter-as segmentation. 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, RFC 7524 requires the use of Inter-area P2MP Segmented Next-Hop Extended Community (S-NH-EC), and the setting of "Leaf Information Required" (LIR) flag in PTA in certain situations. Either of these could be optional in case of EVPN.

Removing these requirements would make the segmentation procedures
transparent to ingress and egress PEs.

RFC 7524 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). That
would allow for more flexible deployment scenarios (e.g. for single-
area provider networks).

This document specifies/clarifies/redefines certain/additional EVPN
BUM procedures, with a salient goal that they're better aligned among
MVPN, EVPN and VPLS. For brevity, only changes/additions to relevant
RFC 7117 and RFC 7524 procedures are specified, instead of repeating
the entire procedures. Note that these are to be applied to EVPN
only, even though sometimes they may sound to be updates to RFC
7117/7524.

2.1. 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 Inter-AS Option-B has to be used, in which 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, provider B uses RSVP-TE P2MP
while provider C uses mLDP). 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 to 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 [RFC 7524] 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 sub-domains.
In this new multicast architecture 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

RFC 7432 defines the format of EVPN NLRI as the following:

```
         +-----------------------------------+
         |     Route Type (1 octet)          |
         +-----------------------------------+
         |       Length (1 octet)            |
         +-----------------------------------+
         | Route Type specific (variable)    |
         +-----------------------------------+
```

So far eight types have been defined:

        + 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 - IGMP Join Synch Route
        + 8 - IGMP 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 Administrative sub-field MUST match
that of the RD in all the EVPN routes from the same advertising
router for a given EVI, except the Leaf A-D route (Section 3.3).

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.

```
+-----------------------------------+
|      RD   (8 octets)              |
+-----------------------------------+
|  Ethernet Tag ID (4 octets)       |
+-----------------------------------+
| Extended Community (8 octets)     |
+-----------------------------------+
```

After Ethernet Tag ID, an Extended Community (EC) is used to identify
the region. Various types and sub-types of ECs provide maximum
flexibility. Note that this is not an EC Attribute, but an 8-octet
field embedded in the NLRI itself, following EC encoding scheme.

3.2. S-PMSI A-D route

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

```
+-----------------------------------+
|      RD   (8 octets)              |
+-----------------------------------+
|  Ethernet Tag ID (4 octets)       |
+-----------------------------------+
| Multicast Source Length (1 octet) |
+-----------------------------------+
|  Multicast Source (Variable)      |
+-----------------------------------+
|  Multicast Group Length (1 octet) |
+-----------------------------------+
|  Multicast Group  (Variable)      |
+-----------------------------------+
|Originator's Addr Length (1 octet) |
+-----------------------------------+
|Originator's Addr (4 or 16 octets) |
+-----------------------------------+
```

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 RFC
7117. The procedures in RFC 7117 also apply (including wildcard
functionality), except that the granularity level is per Ethernet
Tag.

3.3. Leaf-AD route

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

```
          +-----------------------------------+
          |       Route Key (variable)        |
          +-----------------------------------+
          |Originator's Addr Length (1 octet) |
          +-----------------------------------+
          |Originator's Addr (4 or 16 octets) |
          +-----------------------------------+
```

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 "Route Type Specific" field of the route for which this Leaf A-D route is generated.

The general procedures of Leaf A-D route are first specified in RFC 6514 for MVPN. The principles apply to VPLS and EVPN as well. RFC 7117 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 or BIER for all flows. An NVE proxies the IGMP/MLD state that it learns on its ACs to (C-S,C-G) or (C-*,C-G) SMET routes and advertises to other NVEs, and an receiving NVE converts the SMET routes back to IGMP/MLD messages and send 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 nut shell, a source NVE advertises S-SPMSI
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 Requested 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 LIR bit to request Leaf A-D routes, an egress
NVE may omit the Leaf A-D route if it already advertises a
corresponding SMET route, and the source NVE will use that in lieu of
the Leaf A-D route.

5. Inter-AS Segmentation

5.1. Changes to Section 7.2.2 of RFC 7117

The first paragraph of Section 7.2.2.2 of RFC 7117 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, 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
RFC 7117 itself.

The following text at the end of the second bullet:

      "... 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:

> "... If, in order
> to instantiate the segment, the ASBR needs to know the leaves of
> the tree, then the ASBR MUST set the LIR 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:

> "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:

> "If the received Inter-AS A-D route has the LIR flag set in its PTA,
> then a receiving PE must originate a corresponding Leaf A-D route,
> and 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 ASBR that (re-)advertised the
> Inter-AS A-D route then establishes 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 LIR flag in its I-PMSI'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 RFC 7432 explicitly states that the LIR flag must be set to zero.

An implementation of RFC 7432 might have used the Originating
Router's IP Address field of the Inclusive Multicast Ethernet Tag
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 I-PMSI A-D routes, so it
does not have to set the LIR bit set to request Leaf A-D routes. PEs
within the same AS will all have different next hops in their I-PMSI
A-D routes (hence will all be considered as leaves), and PEs from
other ASes will have the next hop in their I-PMSI 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 IBGP
I-PMSI A-D routes, it MUST advertise the same label for all those for
the same Ethernet Tag ID and the same EVI. When an ingress PE builds
its flooding list, multiple routes may have the same (nexthop, label)
tuple and they will 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:

o An ingress PE uses the Next Hop instead of Originating Router's IP
   Address to determine leaves for the I-PMSI tunnel.

o An egress PE sends Leaf A-D routes in response to I-PMSI routes,
   if the PTA has the LIR flag set (by the re-advertising ASBRs).

o 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.

To address this backward compatibility problem, the following
procedure can be used (see Section 6.2 for per-PE/AS/region I-PMSI
A-D routes):

o 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.

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

o The ASBRs in an AS originate per-region I-PMSI A-D routes and
   advertise to their external peers to advertise tunnels used to
   carry traffic from the local AS to other ASes. Depending on the
   types of tunnels being used, the LIR 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.

o 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 AD 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
above procedures have 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 SHOULD include a DF Election EC. The EC and
its use is specified in [I-D.ietf-bess-evpn-df-election-framework].
The AC-DF bit in the DF Election EC SHOULD be cleared. If it is
known that no legacy PEs exist in the AS, the ASBR SHOULD NOT include
the EC and SHOULD 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
[I-D.ietf-bess-evpn-df-election-framework], 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-adveried 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
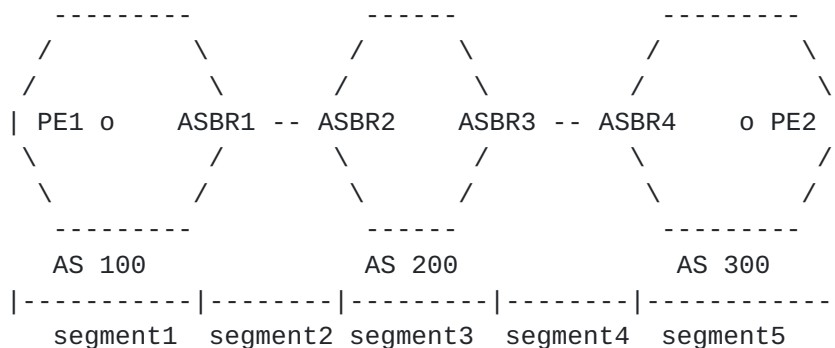NLRI carrying the DF Election EC.

6. Inter-Region Segmentation

6.1. Area vs. Region

   RFC 7524 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.

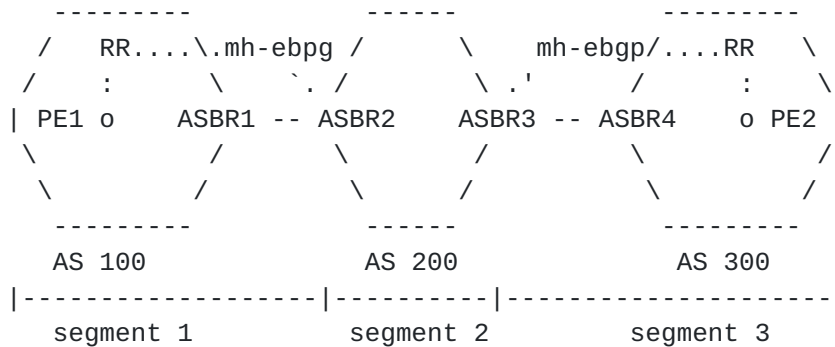   Consider the following example diagram:

```
         ---------           ------              ---------
        /         \         /      \            /         \
       /           \       /        \          /           \
      | PE1 o    ASBR1 -- ASBR2    ASBR3 -- ASBR4    o PE2 |
       \           /       \        /          \           /
        \         /         \      /            \         /
         ---------           ------              ---------
          AS 100             AS 200              AS 300
      |-----------|--------|---------|--------|------------|
         segment1  segment2 segment3  segment4  segment5
```

   The inter-as segmentation procedures specified so far (RFC 6513/6514,
   7117, and Section 5 of this document) requires 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 3. Now the
segment is rooted at ASBR3 and extends across the external link to
PEs.

```
      ---------              ------               ---------
     /   RR....\.mh-ebpg /       \    mh-ebgp/....RR    \
    /    :       \     `. /       \ .'      /       :      \
   | PE1 o     ASBR1 -- ASBR2    ASBR3 -- ASBR4     o PE2 |
    \            /       \          /        \          /
     \          /         \        /          \        /
      ---------              ------               ---------
       AS 100                AS 200                AS 300
     |-------------------|----------|--------------------|
         segment 1         segment 2         segment 3
```

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 route
   key (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 per-region I-PMSI route uses an embedded EC in NLRI to identify a
region. As long as it uniquely identifies the region and the RBRs
for the same region uses the same EC it is permitted. In the case
where an AS number or area ID is needed, the following can be used:

o 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.

o 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.

o For an area ID, a Transitive IPv4-Address-Specific EC of any sub-
   type.

Uses of other particular ECs may be specified in other documents.

6.3. Use of S-NH-EC

RFC 7524 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 AD 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. if a downstream PE/RBR needs to originate Leaf A-D routes, it
simply uses the BGP next hop in the corresponding I/S-PMSI A-D routes
to construct Route Targets.

The advantage of this is that neither ingress nor egress PEs need to
understand/use S-NH-EC, and consistent procedure (based on BGP next
hop) is used for both inter-as and inter-region segmentation.

6.4. Ingress PE's I-PMSI Leaf Tracking

RFC 7524 specifies that when an ingress PE/ASBR (re-)advertises an
VPLS I-PMSI A-D route, it sets the LIR 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 LIR 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

If multi-homing does not span across different ASes or regions,
existing procedures work with segmentation, and a segmentation point
will remove the ESI label from the packets. If an ES is multi-homed
to PEs in different ASes or regions, additional procedures are needed
to work with segmentation. The procedures are well understood but
omitted here until the requirement becomes clear.

8. IANA Considerations

IANA has temporarily assigned the following new EVPN route types:

o 9 - Per-Region I-PMSI A-D route

o 10 - S-PMSI A-D route

o 11 - Leaf A-D route

This document requests IANA to assign one flag bit from the EVPN
Multicast Flags Extended Community:

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

9. Security Considerations

This document does not seem to introduce new security risks, though
this may be revised after further review and scrutiny.

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-df-election-framework]
          Rabadan, J., satyamoh@cisco.com, s., Sajassi, A., Drake,
          J., Nagaraj, K., and S. Sathappan, "Framework for EVPN
          Designated Forwarder Election Extensibility", draft-ietf-
          bess-evpn-df-election-framework-03 (work in progress), May
          2018.

[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.

[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>.

[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>.

   [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>.

12.2. Informative References

   [I-D.ietf-bess-evpn-overlay]
              Sajassi, A., Drake, J., Bitar, N., Shekhar, R., Uttaro,
              J., and W. Henderickx, "A Network Virtualization Overlay
              Solution using EVPN", draft-ietf-bess-evpn-overlay-12
              (work in progress), February 2018.

   [I-D.ietf-bier-architecture]
              Wijnands, I., Rosen, E., Dolganow, A., Przygienda, T., and
              S. Aldrin, "Multicast using Bit Index Explicit
              Replication", draft-ietf-bier-architecture-08 (work in
              progress), September 2017.

   [I-D.zzhang-bier-evpn]
              Zhang, Z., Przygienda, T., Sajassi, A., and J. Rabadan,
              "EVPN BUM Using BIER", draft-zzhang-bier-evpn-00 (work in
              progress), June 2017.

   [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>.

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