BESS Working Group                                    Ali Sajassi
Internet-Draft                                       Samir Thoria
Intended Status: Standards Track                            Cisco

                                                      Keyur Patel
                                                     Derek Yeung
                                                           Arrcus
                                                       John Drake
                                                          Wen Lin
                                                          Juniper

Expires: September 4, 2018                          March 4, 2018


**IGMP and MLD Proxy for EVPN**
**draft-ietf-bess-evpn-igmp-mld-proxy-01**

Abstract

   Ethernet Virtual Private Network (EVPN) solution [RFC 7432] is
   becoming pervasive in data center (DC) applications for Network
   Virtualization Overlay (NVO) and DC interconnect (DCI) services, and
   in service provider (SP) applications for next generation virtual
   private LAN services.

   This draft describes how to support efficiently endpoints running
   IGMP for the above services over an EVPN network by incorporating
   IGMP proxy procedures on EVPN PEs.

Status of this Memo

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

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF), its areas, and its working groups.  Note that
   other groups may also distribute working documents as
   Internet-Drafts.

   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."

   The list of current Internet-Drafts can be accessed at
   http://www.ietf.org/1id-abstracts.html

Table of Contents

## 1  Introduction

Ethernet Virtual Private Network (EVPN) solution [RFC 7432] is
becoming pervasive in data center (DC) applications for Network
Virtualization Overlay (NVO) and DC interconnect (DCI) services, and
in service provider (SP) applications for next generation virtual
private LAN services.

In DC applications, a POD can consist of a collection of servers
supported by several TOR and Spine switches. This collection of
servers and switches are self contained and may have their own
control protocol for intra-POD communication and orchestration.
However, EVPN is used as way of standard inter-POD communication for
both intra-DC and inter-DC. A subnet can span across multiple PODs
and DCs. EVPN provides robust multi-tenant solution with extensive
multi-homing capabilities to stretch a subnet (e.g., VLAN) across
multiple PODs and DCs. There can be many hosts/VMs (e.g., several
hundreds) attached to a subnet that is stretched across several PODs
and DCs.

These hosts/VMs express their interests in multicast groups on a
given subnet/VLAN by sending IGMP membership reports (Joins) for
their interested multicast group(s). Furthermore, an IGMP router
(e.g., IGMPv1) periodically sends membership queries to find out if
there are hosts on that subnet still interested in receiving
multicast traffic for that group. The IGMP/MLD Proxy solution
described in this draft has three objectives to accomplish:

1) Reduce flooding of IGMP messages: just like ARP/ND suppression
mechanism in EVPN to reduce the flooding of ARP messages over EVPN,
it is also desired to have a mechanism to reduce the flood of IGMP
messages (both Queries and Reports) in EVPN.

2) Distributed anycast multicast proxy: it is desired for the EVPN
network to act as a distributed anycast multicast router with respect
to IGMP/MLD proxy function for all the hosts attached to that
subnet.

3) Selective Multicast: to forward multicast traffic over EVPN
network such that it only gets forwarded to the PEs that have
interest in the multicast group(s) - i.e., multicast traffic will not
be forwarded to the PEs that have no receivers attached to them for
that multicast group. This draft shows how this objective may be
achieved when Ingress Replication is used to distribute the multicast
traffic among the PEs.  Procedures for supporting selective multicast
using P2MP tunnels can be found in [bum-procedure-updates]

The first two objectives are achieved by using IGMP/MLD proxy on the

PE and the third objective is achieved by setting up a multicast
tunnel (e.g., ingress replication) only among the PEs that have
interest in that multicast group(s) based on the trigger from
IGMP/MLD proxy processes. The proposed solutions for each of these
objectives are discussed in the following sections.


## 1.1  Terminology

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.

POD: Point of Delivery

ToR: Top of Rack

NV: Network Virtualization

NVO: Network Virtualization Overlay

VNI:  Virtual Network Identifier (for VXLAN)

EVPN: Ethernet Virtual Private Network

IGMP: Internet Group Management Protocol

MLD: Multicast Listener Discovery

EVI: An EVPN instance spanning the Provider Edge (PE) devices
participating in that EVPN

MAC-VRF: A Virtual Routing and Forwarding table for Media Access
Control (MAC) addresses on a PE

Ethernet Segment (ES): When a customer site (device or network) is
connected to one or more PEs via a set of Ethernet links, then that
set of links is referred to as an 'Ethernet segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that
identifies an Ethernet segment is called an 'Ethernet Segment
Identifier'.

PE: Provider Edge device.

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
VLAN-aware bundle service model, an EVI contains multiple BDs. Also,
in this document, BD and subnet are equivalent terms.

Ethernet Tag: An Ethernet tag identifies a particular broadcast
domain, e.g., a VLAN.  An EVPN instance consists of one or more
broadcast domains.

Single-Active Redundancy Mode: When only a single PE, among all the
PEs attached to an Ethernet segment, is allowed to forward traffic
to/from that Ethernet segment for a given VLAN, then the Ethernet
segment is defined to be operating in Single-Active redundancy mode.

All-Active Redundancy Mode: When all PEs attached to an Ethernet
segment are allowed to forward known unicast traffic to/from that
Ethernet segment for a given VLAN, then the Ethernet segment is
defined to be operating in All-Active redundancy mode.

## 2  IGMP Proxy

IGMP Proxy mechanism is used to reduce the flooding of IGMP messages
over EVPN network similar to ARP proxy used in reducing the flooding
of ARP messages over EVPN. It also provides triggering mechanism for
the PEs to setup their underlay multicast tunnels. IGMP Proxy
mechanism consist of two components: a) Proxy for IGMP Reports and b)
Proxy for IGMP Queries.

### 2.1  Proxy Reporting

When IGMP protocol is used between host/VMs and its first hop EVPN
router (EVPN PE), Proxy-reporting is used by the EVPN PE to summarize
(when possible) reports received from downstream hosts and propagate
it in BGP to other PEs that are interested in the info. This is done
by terminating IGMP Reports in the first hop PE, translating and
exchanging the relevant information among EVPN BGP speakers. The
information is again translated back to IGMP message at the recipient
EVPN speaker. Thus it helps create an IGMP overlay subnet using BGP.
In order to facilitate such an overlay, this document also defines a
new EVPN route type NLRI, EVPN Selective Multicast Ethernet Tag
route, along with its procedures to help exchange and register IGMP
multicast groups [section 5].

### 2.1.1  IGMP Membership Report Advertisement in BGP

When a PE wants to advertise an IGMP membership report (Join) using
the BGP EVPN route, it follows the following rules:

1) When the first hop PE receives several IGMP membership reports

(Joins) , belonging to the same IGMP version, from different attached
hosts/VMs for the same (*,G) or (S,G), it only sends a single BGP
message corresponding to the very first IGMP Join. This is because
BGP is a statefull protocol and no further transmission of the same
report is needed. If the IGMP Join is for (*,G), then multicast group
address along with the corresponding version flag (v1, v2, or v3) are
set. In case of IGMPv3, exclude flag also needs to be set to indicate
that no source IP address to be excluded (e.g., include all sources
"*"). If the IGMP Join is for (S,G), then besides setting multicast
group address along with the version flag v3, the source IP address
and the include/exclude flag must be set. It should be noted that
when advertising the EVPN route for (S,G), the only valid version
flag is v3 (i.e., v1 and v2 flags must be set to zero).


2) When the first hop PE receives an IGMPv3 Join for (S,G) on a given
BD, it advertises the corresponding EVPN Selective Multicast Ethernet
Tag (SMET) route regardless of whether the source (S) is attached to
itself or not in order to facilitate the source move in the future.

3) When the first hop PE receives an IGMP version-X Join first for
(*,G) and then later it receives an IGMP version-Y Join for the same
(*,G), then it will re-advertise the same EVPN SMET route with flag
for version-Y set in addition to any previously-set version flag(s).
In other words, the first hop PE does not withdraw the EVPN route
before sending the new route because the flag field is not part of
BGP route key processing.

4) When the first hop PE receives an IGMP version-X Join first for
(*,G) and then later it receives an IGMPv3 Join for the same
multicast group address but for a specific source address S, then the
PE will advertise a new EVPN SMET route with v3 flag set (and v1 and
v2 reset). Include/exclude flag also need to be set accordingly.
Since source IP address is used as part of BGP route key processing,
it is considered as a new BGP route advertisement.


5) When a PE receives an EVPN SMET route with more than one version
flag set, it will generate the corresponding IGMP report for (*,G)
for each version specified in the flag field. With multiple version
flags set, there should be no source IP address in the receive EVPN
route. If there is, then an error should be logged. If v3 flag is set
(in addition to v1 or v2), then the include/exclude flag needs to
indicate "exclude". If not, then an error should be logged. The PE
MUST generate an IGMP membership report (Join) for that (*,G) and
each IGMP version in the version flag.

6) When a PE receives a list of EVPN SMET NLRIs in its BGP update
message, each with a different source IP address and the multicast
group address, and the version flag is set to v3, then the PE
generates an IGMPv3 membership report with a record corresponding to
the list of source IP addresses and the group address along with the
proper indication of inclusion/exclusion.

7) Upon receiving EVPN SMET route(s) and before generating the
corresponding IGMP Join(s), the PE checks to see whether it has any
CE multicast router for that BD on any of its ES's . The PE provides
such check by listening for PIM hellos on that AC (i.e, <ES,BD>). If
it has router's ACs, then the generated IGMP Join(s) are sent to
those ACs. If it doesn't have any router's AC, then no IGMP Join(s)
needs to be generated because sending IGMP Joins to other hosts can
result in unintentionally preventing a host from joining a specific
multicast group for IGMPv1 and IGMPv2 - i.e., if the PE does not
receive a join from the host it will not forward multicast data to
it. Per [RFC4541], when an IGMPv1 or IGMPv2 host receives a
membership report for a group address that it intends to join, the
host will suppress its own membership report for the same group. In
other words, an IGMPv1 or IGMPv2 Join MUST NOT be sent on an AC that
does not lead to a CE multicast router. This message suppression is a
requirement for IGMPv1 and IGMPv2 hosts. This is not a problem for
hosts running IGMPv3 because there is no suppression of IGMP
Membership reports.

## 2.1.1  IGMP Leave Group Advertisement in BGP

When a PE wants to withdraw an EVPN SMET route corresponding to an
IGMPv2 Leave Group (Leave) or IGMPv3 "Leave" equivalent message, it
follows the following rules:

1) For IGMPv1, there is no explicit membership leave; therefore, the
PE needs to periodically send out an IGMP membership query to
determine whether there is any host left who is interested in
receiving traffic directed to this multicast group (this proxy query
function will be described in more details in section 2.2). If there
is no host left, then the PE re-advertises EVPN SMET route with the
v1 version flag reset. If this is the last version flag to be reset,
then instead of re-advertising the EVPN route with all version flags
reset, the PE withdraws the EVPN route for that (*,G).

2) When a PE receives an IGMPv2 Leave Group or its "Leave" equivalent
message for IGMPv3 from its attached host, it checks to see if this
host is the last host who is interested in this multicast group by
sending a query for the multicast group. If the host was indeed the
last one, then the PE re-advertises EVPN SMET Multicast route with
the corresponding version flag reset. If this is the last version

flag to be reset, then instead of re-advertising the EVPN route with
all version flags reset, the PE withdraws the EVPN route for that
(*,G).

3) When a PE receives an EVPN SMET route for a given (*,G), it
compares the received version flags from the route with its per-PE
stored version flags. If the PE finds that a version flag associated
with the (*,G) for the remote PE is reset, then the PE generates IGMP
Leave for that (*,G) toward its local interface (if any) attached to
the multicast router for that multicast group. It should be noted
that the received EVPN route should at least have one version flag
set. If all version flags are reset, it is an error because the PE
should have received an EVPN route withdraw for the last version
flag. If the PE receives an EVPN SMET route withdraw, then it must
remove the remote PE from the OIF list associated with that multicast
group.

4) When a PE receives an EVPN SMET route withdraw, it removes the
remote PE from its OIF list for that multicast group and if there are
no more OIF entries for that multicast group (either locally or
remotely), then the PE MUST stop responding to queries from the
locally attached router (if any). If there is a source for that
multicast group, the PE stops sending multicast traffic for that
source.


## 2.2  Proxy Querier

As mentioned in the previous sections, each PE need to have proxy
querier functionality for the following reasons:

1) To enable the collection of EVPN PEs providing L2VPN service to
act as distributed multicast router with Anycast IP address for all
attached hosts/VMs in that subnet.

2) To enable suppression of IGMP membership reports and queries over
MPLS/IP core.

3) To enable generation of query messages locally to their attached
host. In case of IGMPv1, the PE needs to send out an IGMP membership
query to verify that at least one host on the subnet is still
interested in receiving traffic directed to that group. When there is
no reply to three consecutive IGMP membership queries, the PE times
out the group, stops forwarding multicast traffic to the attached
hosts for that (*,G), and sends a EVPN SMET route associated with
that (*,G) with the version-1 flag reset or withdraws that route.

**3 Operation**

   Consider the EVPN network of figure-1, where there is an EVPN
   instance configured across the PEs shown in this figure (namely PE1,
   PE2, and PE3). Lets consider that this EVPN instance consist of a
   single bridge domain (single subnet) with all the hosts, sources and
   the multicast router shown in this figure connected to this subnet.
   PE1 only has hosts connected to it. PE2 has a mix of hosts and
   multicast source. PE3 has a mix of hosts, multicast source, and
   multicast router. Further more, lets consider that for (S1,G1), R1 is
   used as the multicast router. The following subsections describe the
   IGMP proxy operation in different PEs with regard to whether the
   locally attached devices for that subnet are:

        - only hosts/VMs
        - mix of hosts/VMs and multicast source
        - mix of hosts/VMs, multicast source, and multicast router

```
                          +--------------+
                          |              |
                          |              |
              +----+ |                | +----+
  H1:(*,G1)v1 ---|    | |                | |    |---- H6(*,G1)v2
  H2:(*,G1)v1 ---| PE1| |   IP/MPLS      | | PE2|---- H7(S2,G2)v3
  H3:(*,G1)v2 ---|    | |   Network      | |    |---- S2
  H4:(S2,G2)v3 --|    | |                | |    |
              +----+ |                | +----+
                          |              |
              +----+ |                |
  H5:(S1,G1)v3 --|    | |                |
          S1 ---| PE3| |                |
          R1 ---|    | |                |
              +----+ |                |
                          |              |
                          +--------------+
```
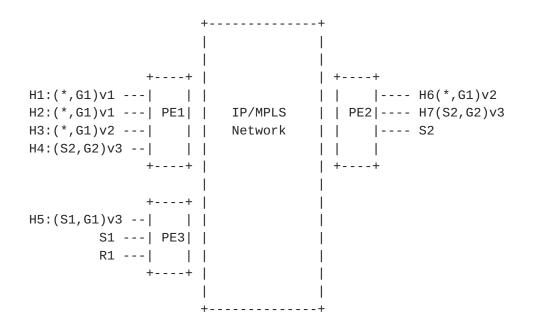
   Figure 1:

**3.1 PE with only attached hosts/VMs for a given subnet**

   When PE1 receives an IGMPv1 Join Report from H1, it does not forward
   this join to any of its other ports (for this subnet) because all
   these local ports are associated with the hosts/VMs. PE1 sends an

EVPN Multicast Group route corresponding to this join for (*,G1) and setting v1 flag. This EVPN route is received by PE2 and PE3 that are the member of the same BD (i.e., same EVI in case of VLAN-based service or <EVI,VLAN> in case of VLAN-aware bundle service). PE3 reconstructs IGMPv1 Join Report from this EVPN BGP route and only sends it to the port(s) with multicast routers attached to it (for that subnet). In this example, PE3 sends the reconstructed IGMPv1 Join Report for (*,G1) to only R1. Furthermore, PE2 although receives the EVPN BGP route, it does not send it to any of its port for that subnet - namely ports associated with H6 and H7.

When PE1 receives the second IGMPv1 Join from H2 for the same multicast group (*,G1), it only adds that port to its OIF list but it doesn't send any EVPN BGP route because there is no change in information. However, when it receives the IGMPv2 Join from H3 for the same (*,G1), besides adding the corresponding port to its OIF list, it re-advertises the previously sent EVPN SMET route with the version-2 flag set.

Finally when PE1 receives the IMGMPv3 Join from H4 for (S2,G2), it advertises a new EVPN SMET route corresponding to it.

## 3.2 PE with mixed of attached hosts/VMs and multicast source

The main difference in here is that when PE2 receives IGMPv3 Join from H7 for (S2,G2), it does not advertises it in BGP because PE2 knows that S2 is attached to its local AC. PE2 adds the port associated with H7 to its OIF list for (S2,G2). The processing for IGMPv2 received from H6 is the same as the v2 Join described in previous section.

## 3.3 PE with mixed of attached hosts/VMs, multicast source and router

The main difference in here relative to the previous two sections is that Join messages received locally needs to be sent to the port associated with router R1. Furthermore, the Joins received via BGP need to be passed to the R1 port but filtered for all other ports.

## 4 All-Active Multi-Homing

Because a CE's LAG flow hashing algorithm is unknown, in an All-Active redundancy mode it must be assumed that the CE can send a given IGMP message to any one of the multi-homed PEs, either DF or non-DF - i.e., different IGMP Join messages can arrive at different PEs in the redundancy group and furthermore their corresponding Leave messages can arrive at PEs that are different from the ones received

the Join messages.  Therefore, all PEs attached to a given ES must
coordinate IGMP Join and Leave Group (x, G) state, where x may be
either '*' or a particular source S, for each BD on that ES. This
allows the DF for that [ES, BD] to correctly advertise or withdraw a
Selective Multicast Ethernet Tag (SMET) route for that (x, G) group
in that BD when needed.

All-Active multihoming PEs for a given ES MUST support IGMP synch
procedures described in this section if they want to perform IGMP
proxy for hosts connects to that ES.

## 4.1 Local IGMP Join Synchronization

When a PE, either DF or non-DF, receives, on a given multihomed ES
operating in All-Active redundancy mode, an IGMP Membership Report
for (x, G), it determines the BD to which the IGMP Membership Report
belongs. If the PE doesn't already have local IGMP Join (x, G) state
for that BD on that ES, it instantiates local IGMP Join (x, G) state
and advertises a BGP IGMP Join Synch route for that [ES, BD]. Local
IGMP Join (x, G) state refers to IGMP Join (x, G) state that is
created as the result of processing an IGMP Membership Report for (x,
G).

The IGMP Join Synch route carries the ES-Import RT for the ES on
which the IGMP Membership Report was received.  Thus it may only go
to the PEs attached to that ES (and not any other PEs).

When a PE, either DF or non-DF, receives an IGMP Join Synch route it
installs that route and if it doesn't already have IGMP Join (x, G)
state for that [ES, BD], it instantiates that IGMP Join (x,G) state -
i.e., IGMP Join (x, G) state is the union of local IGMP Join (x, G)
state and installed IGMP Join Synch route. If the DF is not currently
advertising (originating) a SMET route for that (x, G) group in that
BD, it does so now.

When a PE, either DF or non-DF, deletes its local IGMP Join (x, G)
state for that [ES, BD], it withdraws its BGP IGMP Join Synch route
for that [ES, BD].

When a PE, either DF or non-DF, receives the withdrawal of an IGMP
Join Synch route from another PE it removes that route.  When a PE
has no local IGMP Join (x, G) state and it has no installed IGMP Join
Synch routes, it removes IGMP Join (x, G) state for that [ES, BD].
If the DF no longer has IGMP Join (x, G) state for that BD on any ES
for which it is DF, it withdraws its SMET route for that (x, G) group
in that BD.

I.e., A PE advertises an SMET route for that (x, G) group in that BD

   when it has IGMP Join (x, G) state in that BD on at least one ES for
   which it is DF and it withdraws that SMET route when it does not have
   IGMP Join (x, G) state in that BD on any ES for which it is DF.

## 4.2 Local IGMP Leave Group Synchronization

   When a PE, either DF or non-DF, receives, on a given multihomed ES
   operating in All-Active redundancy mode, an IGMP Leave Group message
   for (x, G) from the attached CE, it determines the BD to which the
   IGMPv2 Leave Group belongs.  Regardless of whether it has IGMP Join
   (x, G) state for that [ES, BD], it initiates the (x, G) leave group
   synchronization procedure, which consists of the following steps:

   1) It computes the Maximum Response Time, which is the duration of
   (x, G) leave group synchronization procedure.  This is the product of
   two locally configured values, Last Member Query Count and Last
   Member Query Interval (described in Section 3 of [RFC2236]), plus
   delta, the time it takes for a BGP advertisement to propagate between
   the PEs attached to the multihomed ES (delta is a consistently
   configured value on all PEs attached to the multihomed ES).

   2) It starts the Maximum Response Time timer. Note that the receipt
   of subsequent IGMP Leave Group messages or BGP Leave Synch routes for
   (x, G) do not change the value of a currently running Maximum
   Response Time timer and are ignored by the PE.

   3) It initiates the Last Member Query procedure described in Section
   3 of [RFC2236]; viz, it sends a number of Group-Specific Query (x, G)
   messages (Last Member Query Count) at a fixed interval (Last Member
   Query Interval) to the attached CE.

   4) It advertises an IGMP Leave Synch route for that that [ES, BD].
   This route notifies the other multihomed PEs attached to the given
   multihomed ES that it has initiated an (x, G) leave group
   synchronization procedure; i.e., it carries the ES-Import RT for the
   ES on which the IGMP Leave Group was received.  It also contains the
   Maximum Response Time and the Leave Group Synchronization Procedure
   Sequence number. The latter identifies the specific (x, G) leave
   group synchronization procedure initiated by the advertising PE,
   which increments the value whenever it initiates a procedure.

   5) When the Maximum Response Timer expires, the PE that has
   advertised the IGMP Leave Synch route withdraws it.


## 4.2.1 Remote Leave Group Synchronization

   When a PE, either DF or non-DF, receives an IGMP Leave Synch route it

installs that route and it starts a timer for (x, G) on the specified
[ES, BD] whose value is set to the Maximum Response Time in the
received IGMP Leave Synch route.  Note that the receipt of subsequent
IGMPv2 Leave Group messages or BGP Leave Synch routes for (x, G) do
not change the value of a currently running Maximum Response Time
timer and are ignored by the PE.

## 4.2.2 Common Leave Group Synchronization

If a PE attached to the multihomed ES receives an IGMP Membership
Report for (x, G) before the Maximum Response Time timer expires, it
advertises a BGP IGMP Join Synch route for that [ES, BD]. If it
doesn't already have local IGMP Join (x, G) state for that [ES, BD],
it instantiates local IGMP Join (x, G) state. If the DF is not
currently advertising (originating) a SMET route for that (x, G)
group in that BD, it does so now.

If a PE attached to the multihomed ES receives an IGMP Join Synch
route for (x, G) before the Maximum Response Time timer expires, it
installs that route and if it doesn't already have IGMP Join (x, G)
state for that BD on that ES, it instantiates that IGMP Join (x,G)
state. If the DF is not currently advertising (originating) a SMET
route for that (x, G) group in that BD, it does so now.

When the Maximum Response Timer expires a PE that has advertised an
IGMP Leave Synch route, withdraws it.  Any PE attached to the
multihomed ES, that started the Maximum Response Time and has no
local IGMP Join (x, G) state and no installed IGMP Join Synch routes,
it removes IGMP Join (x, G) state for that [ES, BD].  If the DF no
longer has IGMP Join (x, G) state for that BD on any ES for which it
is DF, it withdraws its SMET route for that (x, G) group in that BD.

## 5 Single-Active Multi-Homing

Note that to facilitate state synchronization after failover, the PEs
attached to a multihomed ES operating in Single-Active redundancy
mode should also coordinate IGMP Join (x, G) state.  In this case all
IGMP Join messages are received by the DF and distributed to the non-
DF PEs using the procedures described above.

## 6 Selective Multicast Procedures for IR tunnels

If an ingress PE uses ingress replication, then for a given (x, G)
group in a given BD:

1)  It sends (x, G) traffic to the set of PEs not supporting IGMP

Proxy.  This set consists of any PE that has advertised an Inclusive
Multicast Tag route for the BD without the "IGMP Proxy Support" flag.

2)  It sends (x, G) traffic to the set of PEs supporting IGMP Proxy
and having listeners for that (x, G) group in that BD. This set
consists of any PE that has advertised an Inclusive Multicast Tag
route for the BD with the "IGMP Proxy Support" flag and that has
advertised an SMET route for that (x, G) group in that BD.

If an ingress PE's Selective P-Tunnel for a given BD uses P2MP and
all of the PEs in the BD support that tunnel type and IGMP, then for
a given (x, G) group in a given BD it sends (x, G) traffic using the
Selective P-Tunnel for that (x, G) group in that BD.  This tunnel
will include those PEs that have advertised an SMET route for that
(x, G) group on that BD (for Selective P-tunnel) but it may include
other PEs as well (for Aggregate Selective P-tunnel).

## 7 BGP Encoding

This document defines three new BGP EVPN routes to carry IGMP
membership reports. This route type is known as:

         + 6 -  Selective Multicast Ethernet Tag Route
         + 7 -  IGMP Join Synch Route
         + 8 - IGMP Leave Synch Route

The detailed encoding and procedures for this route type is described
in subsequent section.

## 7.1 Selective Multicast Ethernet Tag Route

An Selective Multicast Ethernet Tag route type specific EVPN NLRI
consists of the following:

```
                    +---------------------------------------+
                    |  RD (8 octets)                        |
                    +---------------------------------------+
                    |  Ethernet Tag ID (4 octets)           |
                    +---------------------------------------+
                    |  Multicast Source Length (1 octet)    |
                    +---------------------------------------+
                    |  Multicast Source Address (variable)  |
                    +---------------------------------------+
                    |  Multicast Group Length (1 octet)     |
                    +---------------------------------------+
                    |  Multicast Group Address (Variable)   |
                    +---------------------------------------+
                    |  Originator Router Length (1 octet)   |
                    +---------------------------------------+
                    |  Originator Router Address (variable) |
                    +---------------------------------------+
                    |  Flags (1 octets) (optional)          |
                    +---------------------------------------+
```

For the purpose of BGP route key processing, all the fields are
considered to be part of the prefix in the NLRI except for the one-
octet optional flag field (if included). The Flags fields are defined
as follows:

```
                 0  1  2  3  4  5  6  7
               +--+--+--+--+--+--+--+--+
               | reserved  |IE|v3|v2|v1|
               +--+--+--+--+--+--+--+--+
```

The least significant bit, bit 7 indicates support for IGMP version
1.

The second least significant bit, bit 6 indicates support for IGMP
version 2.

The third least significant bit, bit 5 indicates support for IGMP
version 3.

The forth least significant bit, bit 4 indicates whether the (S, G)
information carried within the route-type is of Include Group type
(bit value 0) or an Exclude Group type (bit value 1). The Exclude
Group type bit MUST be ignored if bit 5 is not set.

This EVPN route type is used to carry tenant IGMP multicast group
information. The flag field assists in distributing IGMP membership
interest of a given host/VM for a given multicast route. The version

   bits help associate IGMP version of receivers participating within
   the EVPN domain.

   The include/exclude bit helps in creating filters for a given
   multicast route.


**7.1.1 Constructing the Selective Multicast Ethernet Tag route**

   This section describes the procedures used to construct the Selective
   Multicast Ethernet Tag (SMET) route. Support for this route type is
   optional.

   The Route Distinguisher (RD) SHOULD be a Type 1 RD [RFC4364].  The
   value field comprises an IP address of the PE (typically, the
   loopback address) followed by a number unique to the PE.

   The Ethernet Tag ID MUST be set as follows:

   EVI is VLAN-Based or VLAN Bundle service - set to 0
   EVI is VLAN-Aware Bundle service without translation - set to
   the customer VID for that BD
   EVI is VLAN-Aware Bundle service with translation - set to the
   normalized Ethernet Tag ID - e.g., normalized VID


   The Multicast Source length MUST be set to length of multicast source
   address in bits. In case of a (*, G) Join, the Multicast Source
   Length is set to 0.

   The Multicast Source is the Source IP address of the IGMP membership
   report. In case of a (*, G) Join, this field does not exist.

   The Multicast Group length MUST be set to length of multicast group
   address in bits.

   The Multicast Group is the Group address of the IGMP membership
   report.

   The Originator Router Length is the length of the Originator Router
   address in bits.

   The Originator Router Address is the IP address of Router Originating
   the prefix. It should be noted that using the "Originating Router's
   IP address" field is needed for local-bias procedures and may be
   needed for building inter-AS multicast underlay tunnels where BGP
   next hop can get over written.

The Flags field indicates the version of IGMP protocol from which the membership report was received. It also indicates whether the multicast group had INCLUDE or EXCLUDE bit set.

IGMP protocol is used to receive group membership information from hosts/VMs by TORs. Upon receiving the hosts/VMs expression of interest of a particular group membership, this information is then forwarded using Ethernet Multicast Source Group Route NLRI. The NLRI also keeps track of receiver's IGMP protocol version and any "source filtering" for a given group membership. All EVPN SMET routes are announced with per-EVI Route Target extended communities.

## 7.2  IGMP Join Synch Route

This EVPN route type is used to coordinate IGMP Join (x,G) state for a given BD between the PEs attached to a given ES operating in All-Active (or Single-Active) redundancy mode and it consists of following:

```
        +---------------------------------------------------+
        |  RD (8 octets)                                    |
        +---------------------------------------------------+
        | Ethernet Segment Identifier (10 octets)           |
        +---------------------------------------------------+
        |  Ethernet Tag ID  (4 octets)                      |
        +---------------------------------------------------+
        |  Multicast Source Length (1 octet)                |
        +---------------------------------------------------+
        |  Multicast Source Address (variable)              |
        +---------------------------------------------------+
        |  Multicast Group Length (1 octet)                 |
        +---------------------------------------------------+
        |  Multicast Group Address (Variable)               |
        +---------------------------------------------------+
        |  Originator Router Length (1 octet)               |
        +---------------------------------------------------+
        |  Originator Router Address (variable)             |
        +---------------------------------------------------+
        |  Flags (1 octet)                                  |
        +---------------------------------------------------+
```

For the purpose of BGP route key processing, all the fields are considered to be part of the prefix in the NLRI except for the one-octet Flags field, whose fields are defined as follows:

```
              0  1  2  3  4  5  6  7
            +--+--+--+--+--+--+--+--+
            | reserved  |IE|v3|v2|v1|
            +--+--+--+--+--+--+--+--+
```

The least significant bit, bit 7 indicates support for IGMP version
1.  The second least significant bit, bit 6 indicates support for
IGMP version 2.     The third least significant bit, bit 5 indicates
support for IGMP version 3.  The fourth least significant bit, bit 4
indicates whether the (S, G) information carried within the route-
type is of Include Group type (bit value 0) or an Exclude Group type
(bit value 1). The Exclude Group type bit MUST be ignored if bit 5 is
not set.

The Flags field assists in distributing IGMP membership interest of a
given host/VM for a given multicast route. The version bits help
associate IGMP version of receivers participating within the EVPN
domain.  The include/exclude bit helps in creating filters for a
given multicast route.

### 7.2.1  Constructing the IGMP Join Synch Route

This section describes the procedures used to construct the IGMP Join
Synch route.  Support for this route type is optional. If a PE does
not support this route, then it MUST not indicate that it supports
'IGMP proxy' in Multicast Flag extended community for the EVIs
corresponding to its multi-homed Ethernet Segments. An IGMP Join
Synch route is advertised with an ES-Import Route Target extended
community whose value is set to the ESI for the ES on which the IGMP
Join was received.

The Route Distinguisher (RD) SHOULD be a Type 1 RD [RFC4364].  The
value field comprises an IP address of the PE (typically, the
loopback address) followed by a number unique to the PE.

The Ethernet Segment Identifier (ESI) MUST be set to the 10-octet
value defined for the ES.

The Ethernet Tag ID MUST be set as follows:

EVI is VLAN-Based or VLAN Bundle service - set to 0
EVI is VLAN-Aware Bundle service without translation - set to
the customer VID for the BD
EVI is VLAN-Aware Bundle service with translation - set to the
normalized Ethernet Tag ID - e.g., normalized VID

   The Multicast Source length MUST be set to length of multicast source
   address in bits. In case of a (*, G) Join, the Multicast Source
   Length is set to 0.

   The Multicast Source is the Source IP address of the IGMP membership
   report.  In case of a (*, G) Join, this field does not exist.

   The Multicast Group length MUST be set to length of multicast group
   address in bits.

   The Multicast Group is the Group address of the IGMP membership
   report.

   The Originator Router Length is the length of the Originator Router
   address in bits.

   The Originator Router Address is the IP address of Router Originating
   the prefix.

   The Flags field indicates the version of IGMP protocol from which the
   membership report was received. It also indicates whether the
   multicast group had INCLUDE or EXCLUDE bit set.


**[7.3](#) IGMP Leave Synch Route This EVPN route type is used to coordinate
   IGMP Leave Group (x,G) state for a given BD between the PEs attached**
   to a given ES operating in All-Active (or Single-Active) redundancy
   mode and it consists of following:

```
+---------------------------------------------------+
|  RD (8 octets)                                    |
+---------------------------------------------------+
| Ethernet Segment Identifier (10 octets)           |
+---------------------------------------------------+
|  Ethernet Tag ID  (4 octets)                      |
+---------------------------------------------------+
|  Multicast Source Length (1 octet)                |
+---------------------------------------------------+
|  Multicast Source Address (variable)              |
+---------------------------------------------------+
|  Multicast Group Length (1 octet)                 |
+---------------------------------------------------+
|  Multicast Group Address (Variable)               |
+---------------------------------------------------+
|  Originator Router Length (1 octet)               |
+---------------------------------------------------+
|  Originator Router Address (variable)             |
+---------------------------------------------------+
|  Leave Group Synchronization  # (4 octets)        |
+---------------------------------------------------+
|  Maximum Response Time (1 octet)                  |
+---------------------------------------------------+
|  Flags (1 octet)                                  |
+---------------------------------------------------+
```

For the purpose of BGP route key processing, all the fields are
considered to be part of the prefix in the NLRI except for the
Maximum Response Time and the one-octet Flags field, whose fields are
defined as follows:

```
          0  1  2  3  4  5  6  7
        +--+--+--+--+--+--+--+--+
        | reserved  |IE|v3|v2|v1|
        +--+--+--+--+--+--+--+--+
```

The least significant bit, bit 7 indicates support for IGMP version
1.  The second least significant bit, bit 6 indicates support for
IGMP version 2.    The third least significant bit, bit 5 indicates
support for IGMP version 3.  The fourth least significant bit, bit 4
indicates whether the (S, G) information carried within the route-
type is of Include Group type (bit value 0) or an Exclude Group type
(bit value 1). The Exclude Group type bit MUST be ignored if bit 5 is
not set.

The Flags field assists in distributing IGMP membership interest of a
given host/VM for a given multicast route. The version bits help
associate IGMP version of receivers participating within the EVPN
domain.  The include/exclude bit helps in creating filters for a
given multicast route.


### 7.3.1  Constructing the IGMP Leave Synch Route

This section describes the procedures used to construct the IGMP Join
Synch route.  Support for this route type is optional. If a PE does
not support this route, then it MUST not indicate that it supports
'IGMP proxy' in Multicast Flag extended community for the EVIs
corresponding to its multi-homed Ethernet Segments. An IGMP Join
Synch route is advertised with an ES-Import Route Target extended
community whose value is set to the ESI for the ES on which the IGMP
Join was received.

The Route Distinguisher (RD) SHOULD be a Type 1 RD [RFC4364].  The
value field comprises an IP address of the PE (typically, the
loopback address) followed by a number unique to the PE.

The Ethernet Segment Identifier (ESI) MUST be set to the 10-octet
value defined for the ES.


The Ethernet Tag ID MUST be set as follows:

EVI is VLAN-Based or VLAN Bundle service - set to 0
EVI is VLAN-Aware Bundle service without translation - set to
the customer VID for the BD
EVI is VLAN-Aware Bundle service with translation - set to the
normalized Ethernet Tag ID - e.g., normalized VID


The Multicast Source length MUST be set to length of multicast source
address in bits. In case of a (*, G) Join, the Multicast Source
Length is set to 0.

The Multicast Source is the Source IP address of the IGMP membership
report.  In case of a (*, G) Join, this field does not exist.

The Multicast Group length MUST be set to length of multicast group
address in bits.

The Multicast Group is the Group address of the IGMP membership
report.

The Originator Router Length is the length of the Originator Router
address in bits.

The Originator Router Address is the IP address of Router Originating
the prefix.

The Flags field indicates the version of IGMP protocol from which the
membership report was received. It also indicates whether the
multicast group had INCLUDE or EXCLUDE bit set.


## 7.4 Multicast Flags Extended Community

The 'Multicast Flags' extended community is a new EVPN extended
community.  EVPN extended communities are transitive extended
communities with a Type field value of 6.  IANA will assign a Sub-
Type from the 'EVPN Extended Community Sub-Types' registry.

A PE that supports IGMP proxy on a given BD MUST attach this extended
community to the  Inclusive Multicast Ethernet Tag (IMET) route it
advertises for that BD and it Must set the IGMP Proxy Support flag to
1. Note that an [RFC7432] compliant PE will not advertise this
extended community so its absence indicates that the advertising PE
does not support IGMP Proxy.

The advertisement of this extended community enables more efficient
multicast tunnel setup from the source PE specially for ingress
replication - i.e., if an egress PE supports IGMP proxy but doesn't
have any interest in a given (x, G), it advertises its IGMP proxy
capability using this extended community but it does not advertise
any SMET route for that (x, G). When the source PE (ingress PE)
receives such advertisements from the egress PE, it does not
replicate the multicast traffic to that egress PE; however, it does
replicate the multicast traffic to the egress PEs that don't
advertise such capability even if they don't have any interests in
that (x, G).

A Multicast Flags extended community is encoded as an 8-octet value,
as follows:

```
                        1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Type=0x06   | Sub-Type=TBD |        Flags (2 Octets)         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                         Reserved=0                           |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The low-order bit of the Flags is defined as the "IGMP Proxy Support"
bit.  A value of 1 means that the PE supports IGMP Proxy as defined
in this document, and a value of 0 means that the PE does not support
IGMP proxy. The absence of this extended community also means that
the PE doesn't not support IGMP proxy.

### 7.5  EVI-RT Extended Community

The 'EVI-RT' extended community is a new EVPN extended community.
EVPN extended communities are transitive extended communities with a
Type field value of 6.  IANA will assign a Sub-Type from the 'EVPN
Extended Community Sub-Types' registry.

A PE that supports IGMP synch procedures for All-Active (or Single-
Active) multi-homed ES, MUST attach this extended community to either
IGMP Join Synch route (sec 7.2) or IGMP Leave Synch route (sec 7.3).
This extended community carries the RT associated with the EVI so
that the receiving PE can identify the EVI properly. The reason
standard format RT is not used, is to avoid distribution of these
routes beyond the group of multihoming PEs for that ES.

```
                        1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Type=0x06   | Sub-Type=TBD |       RT associated with EVI    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |            RT associated with the EVI  (cont.)               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

### 8  Acknowledgement

### 9  Security Considerations

Same security considerations as [RFC7432].

## 10  IANA Considerations

IANA has allocated the following EVPN Extended Community sub-types in [RFC7153], and this document is the only reference for them.

0x09    Multicast Flags Extended Community   [this document]  0x0A
EVI-RT Extended Community           [this document]

This document requests the following EVPN route types from IANA registry.

    + 6 - Selective Multicast Ethernet Tag Route    + 7 - IGMP Join
Synch Route     + 8 - IGMP Leave Synch Route

## 11  References

## 11.1  Normative References

[KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate
           Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4360] S. Sangli et al, ""BGP Extended Communities Attribute",
           February, 2006.

[RFC7432] Sajassi et al., "BGP MPLS Based Ethernet VPN", February,
           2015.

## 11.2  Informative References

[ETREE-FMWK] Key et al., "A Framework for E-Tree Service over MPLS
Network", draft-ietf-l2vpn-etree-frwk-03, work in progress, September
2013.

[PBB-EVPN] Sajassi et al., "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-
05.txt, work in progress, October, 2013.

[RFC4541] Christensen, M., Kimball, K., and F. Solensky,
"Considerations for IGMP and MLD snooping PEs", RFC 4541, 2006.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Samir Thoria
Cisco
Email: sthoria@cisco.com


Keyur Patel
Arrcus
Email: keyur@arrcus.com


Derek Yeung
Arrcus
Email: derek@arrcus.com


John Drake
Juniper
Email: jdrake@juniper.net


Wen Lin
Juniper
Email: wlin@juniper.net