BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: May 6, 2021

J. Rabadan, Ed.
J. Kotalwar
S. Sathappan
Nokia
Z. Zhang
W. Lin
Juniper
E. Rosen
Individual

November 2, 2020

Multicast Source Redundancy in EVPN Networks
draft-skr-bess-evpn-redundant-mcast-source-02

Abstract

EVPN supports intra and inter-subnet IP multicast forwarding. However, EVPN (or conventional IP multicast techniques for that matter) do not have a solution for the case where: a) a given multicast group carries more than one flow (i.e., more than one source), and b) it is desired that each receiver gets only one of the several flows. Existing multicast techniques assume there are no redundant sources sending the same flow to the same IP multicast group, and, in case there were redundant sources, the receiver's application would deal with the received duplicated packets. This document extends the existing EVPN specifications and assumes that IP Multicast source redundancy may exist. It also assumes that, in case two or more sources send the same IP Multicast flows into the tenant domain, the EVPN PEs need to avoid that the receivers get packet duplication by following the described procedures.

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 6, 2021.

Table of Contents

1. Introduction 2
1.1. Terminology 4
1.2. Background on IP Multicast Delivery in EVPN Networks . . 6
1.2.1. Intra-subnet IP Multicast Forwarding 6
1.2.2. Inter-subnet IP Multicast Forwarding 7
1.3. Multi-Homed IP Multicast Sources in EVPN 8
1.4. The Need for Redundant IP Multicast Sources in EVPN . . . 10
2. Solution Overview 10
3. BGP EVPN Extensions 12
4. Warm Standby (WS) Solution for Redundant G-Sources 13
4.1. WS Example in an OISM Network 15
4.2. WS Example in a Single-BD Tenant Network 17
5. Hot Standby (HS) Solution for Redundant G-Sources 18
5.1. Use of BFD in the HS Solution 21
5.2. HS Example in an OISM Network 22
5.3. HS Example in a Single-BD Tenant Network 26
6. Security Considerations 26
7. IANA Considerations 26
8. References 26
8.1. Normative References 26
8.2. Informative References 27
Appendix A. Acknowledgments 28
Appendix B. Contributors 28
Authors' Addresses 28

1. Introduction

Intra and Inter-subnet IP Multicast forwarding are supported in EVPN networks. [I-D.ietf-bess-evpn-igmp-mld-proxy] describes the procedures required to optimize the delivery of IP Multicast flows when Sources and Receivers are connected to the same EVPN BD

(Broadcast Domain), whereas [I-D.ietf-bess-evpn-irb-mcast] specifies the procedures to support Inter-subnet IP Multicast in a tenant network. Inter-subnet IP Multicast means that IP Multicast Source and Receivers of the same multicast flow are connected to different BDs of the same tenant.

[I-D.ietf-bess-evpn-igmp-mld-proxy], [I-D.ietf-bess-evpn-irb-mcast] or conventional IP multicast techniques do not have a solution for the case where a given multicast group carries more than one flow (i.e., more than one source) and it is desired that each receiver gets only one of the several flows. Multicast techniques assume there are no redundant sources sending the same flows to the same IP multicast group, and, in case there were redundant sources, the receiver's application would deal with the received duplicated packets.

As a workaround in conventional IP multicast (PIM or MVPN networks), if all the redundant sources are given the same IP address, each receiver will get only one flow. The reason is that, in conventional IP multicast, (S,G) state is always created by the RP (Rendezvous Point), and sometimes by the Last Hop Router (LHR). The (S,G) state always binds the (S,G) flow to a source-specific tree, rooted at the source IP address. If multiple sources have the same IP address, one may end up with multiple (S,G) trees. However, the way the trees are constructed ensures that any given LHR or RP is on at most one of them. The use of an anycast address assigned to multiple sources may be useful for warm standby redundancy solutions. However, on one hand, it's not really helpful for hot standby redundancy solutions and on the other hand, configuring the same IP address (in particular IPv4 address) in multiple sources may bring issues if the sources need to be reached by IP unicast traffic or if the sources are attached to the same Broadcast Domain.

In addition, in the scenario where several G-sources are attached via EVPN/OISM, there is not necessarily any (S,G) state created for the redundant sources. The LHRs may have only (*,G) state, and there may not be an RP (creating (S,G) state) either. Therefore, this document extends the above two specifications and assumes that IP Multicast source redundancy may exist. It also assumes that, in case two or more sources send the same IP Multicast flows into the tenant domain, the EVPN PEs need to avoid that the receivers get packet duplication.

The solution provides support for Warm Standby (WS) and Hot Standby (HS) redundancy. WS is defined as the redundancy scenario in which the upstream PEs attached to the redundant sources of the same tenant, make sure that only one source of the same flow can send multicast to the interested downstream PEs at the same time. In HS the upstream PEs forward the redundant multicast flows to the

downstream PEs, and the downstream PEs make sure only one flow is forwarded to the interested attached receivers.

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.

- PIM: Protocol Independent Multicast.
- MVPN: Multicast Virtual Private Networks.
- OISM: Optimized Inter-Subnet Multicast, as in [I-D.ietf-bess-evpn-irb-mcast].
- Broadcast Domain (BD): an emulated ethernet, such that two systems on the same BD will receive each other's link-local broadcasts. In this document, BD also refers to the instantiation of a Broadcast Domain on an EVPN PE. An EVPN PE can be attached to one or multiple BDs of the same tenant.
- Designated Forwarder (DF): as defined in [RFC7432], an ethernet segment may be multi-homed (attached to more than one PE). An ethernet segment may also contain multiple BDs, of one or more EVIs. For each such EVI, one of the PEs attached to the segment becomes that EVI's DF for that segment. Since a BD may belong to only one EVI, we can speak unambiguously of the BD's DF for a given segment.
- Upstream PE: in this document an Upstream PE is referred to as the EVPN PE that is connected to the IP Multicast source or closest to it. It receives the IP Multicast flows on local ACs (Attachment Circuits).
- Downstream PE: in this document a Downstream PE is referred to as the EVPN PE that is connected to the IP Multicast receivers and gets the IP Multicast flows from remote EVPN PEs.
- G-traffic: any frame with an IP payload whose IP Destination Address (IP DA) is a multicast group G.
- G-source: any system sourcing IP multicast traffic to G.
- SFG: Single Flow Group, i.e., a multicast group address G which represents traffic that contains only a single flow. However,

multiple sources – with the same or different IP – may be transmitting an SFG.

- Redundant G-source: a host or router that transmits an SFG in a tenant network where there are more hosts or routers transmitting the same SFG. Redundant G-sources for the same SFG SHOULD have different IP addresses, although they MAY have the same IP address when in different BDs of the same tenant network. Redundant G-sources are assumed NOT to be "bursty" in this document (typical example are Broadcast TV G-sources or similar).

- P-tunnel: Provider tunnel refers to the type of tree a given upstream EVPN PE uses to forward multicast traffic to downstream PEs. Examples of P-tunnels supported in this document are Ingress Replication (IR), Assisted Replication (AR), Bit Indexed Explicit Replication (BIER), multicast Label Distribution Protocol (mLDP) or Point to Multi-Point Resource Reservation protocol with Traffic Engineering extensions (P2MP RSVP-TE).

- Inclusive Multicast Tree or Inclusive Provider Multicast Service Interface (I-PMSI): defined in [RFC6513], in this document it is applicable only to EVPN and refers to the default multicast tree for a given BD. All the EVPN PEs that are attached to a specific BD belong to the I-PMSI for the BD. The I-PMSI trees are signaled by EVPN Inclusive Multicast Ethernet Tag (IMET) routes.

- Selective Multicast Tree or Selective Provider Multicast Service Interface (S-PMSI): defined in [RFC6513], in this document it is applicable only to EVPN and refers to the multicast tree to which only the interested PEs of a given BD belong to. There are two types of EVPN S-PMSIs:

  * EVPN S-PMSIs that require the advertisement of S-PMSI AD routes from the upstream PE, as in [EVPN-BUM]. The interested downstream PEs join the S-PMSI tree as in [EVPN-BUM].

  * EVPN S-PMSIs that don’t require the advertisement of S-PMSI AD routes. They use the forwarding information of the IMET routes, but upstream PEs send IP Multicast flows only to downstream PEs issuing Selective Multicast Ethernet Tag (SMET) routes for the flow. These S-PMSIs are only supported with the following P-tunnels: Ingress Replication (IR), Assisted Replication (AR) and BIER.

This document also assumes familiarity with the terminology of [RFC7432], [RFC4364], [RFC6513], [RFC6514], [I-D.ietf-bess-evpn-igmp-mld-proxy], [I-D.ietf-bess-evpn-irb-mcast], [EVPN-RT5] and [EVPN-BUM].

1.2. Background on IP Multicast Delivery in EVPN Networks

IP Multicast is all about forwarding a single copy of a packet from a source S to a group of receivers G along a multicast tree. That multicast tree can be created in an EVPN tenant domain where S and the receivers for G are connected to the same BD or different BD. In the former case, we refer to Intra-subnet IP Multicast forwarding, whereas the latter case will be referred to as Inter-subnet IP Multicast forwarding.

1.2.1. Intra-subnet IP Multicast Forwarding

When the source S1 and receivers interested in G1 are attached to the same BD, the EVPN network can deliver the IP Multicast traffic to the receivers in two different ways (Figure 1):

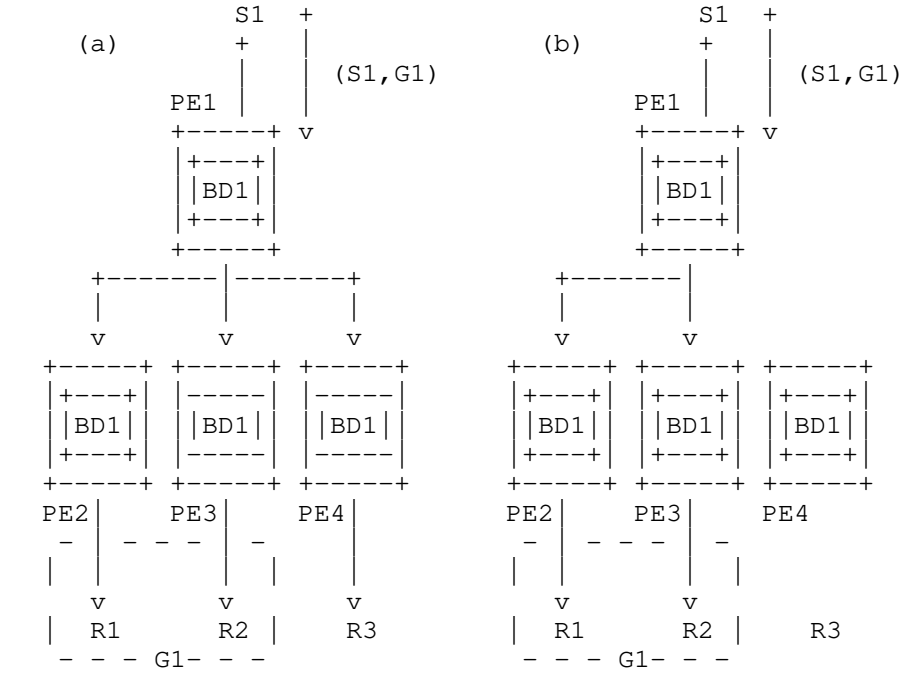

Figure 1: Intra-subnet IP Multicast

Model (a) illustrated in Figure 1 is referred to as "IP Multicast delivery as BUM traffic". This way of delivering IP Multicast traffic does not require any extensions to [RFC7432], however, it sends the IP Multicast flows to non-interested receivers, such as e.g., R3 in Figure 1. In this example, downstream PEs can snoop IGMP/MLD messages from the receivers so that layer-2 multicast state

is created and, for instance, PE4 can avoid sending (S1,G1) to R3, since R3 is not interested in (S1,G1).

Model (b) in Figure 1 uses an S-PMSI to optimize the delivery of the (S1,G1) flow. For instance, assuming PE1 uses IR, PE1 sends (S1,G1) only to the downstream PEs that issued an SMET route for (S1,G1), that is, PE2 and PE3. In case PE1 uses any P-tunnel different than IR, AR or BIER, PE1 will advertise an S-PMSI A-D route for (S1,G1) and PE2/PE2 will join that tree.

Procedures for Model (b) are specified in [I-D.ietf-bess-evpn-igmp-mld-proxy].

1.2.2. Inter-subnet IP Multicast Forwarding

If the source and receivers are attached to different BDs of the same tenant domain, the EVPN network can also use Inclusive or Selective Trees as depicted in Figure 2, models (a) and (b) respectively.

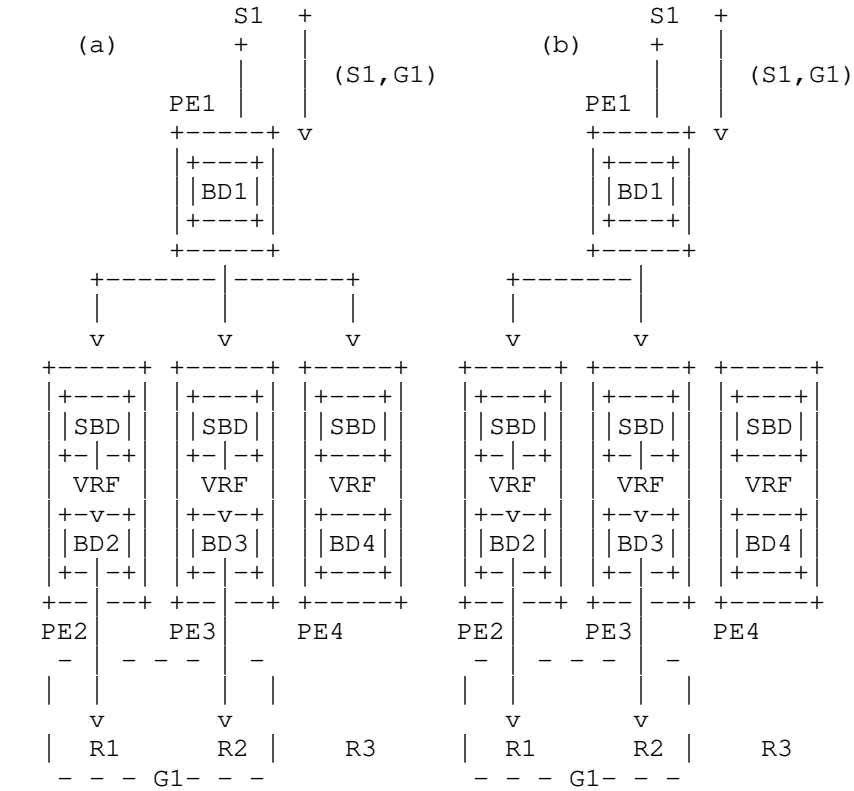

Figure 2: Inter-subnet IP Multicast

[I-D.ietf-bess-evpn-irb-mcast] specifies the procedures to optimize the Inter-subnet Multicast forwarding in an EVPN network. The IP Multicast flows are always sent in the context of the source BD. As described in [I-D.ietf-bess-evpn-irb-mcast], if the downstream PE is not attached to the source BD, the IP Multicast flow is received on the SBD (Supplementary Broadcast Domain), as in the example in Figure 2.

[I-D.ietf-bess-evpn-irb-mcast] supports Inclusive or Selective Multicast Trees, and as explained in Section 1.2.1, the Selective Multicast Trees are setup in a different way, depending on the P-tunnel being used by the source BD. As an example, model (a) in Figure 2 illustrates the use of an Inclusive Multicast Tree for BD1 on PE1. Since the downstream PEs are not attached to BD1, they will all receive (S1,G1) in the context of the SBD and will locally route the flow to the local ACs. Model (b) uses a similar forwarding model, however PE1 sends the (S1,G1) flow in a Selective Multicast Tree. If the P-tunnel is IR, AR or BIER, PE1 does not need to advertise an S-PMSI A-D route.

[I-D.ietf-bess-evpn-irb-mcast] is a superset of the procedures in [I-D.ietf-bess-evpn-igmp-mld-proxy], in which sources and receivers can be in the same or different BD of the same tenant. [I-D.ietf-bess-evpn-irb-mcast] ensures every upstream PE attached to a source will learn of all other PEs (attached to the same Tenant Domain) that have interest in a particular set of flows. This is because the downstream PEs advertise SMET routes for a set of flows with the SBD's Route Target and they are imported by all the Upstream PEs of the tenant. As a result of that, inter-subnet multicasting can be done within the Tenant Domain, without requiring any Rendezvous Points (RP), shared trees, UMH selection or any other complex aspects of conventional multicast routing techniques.

1.3. Multi-Homed IP Multicast Sources in EVPN

Contrary to conventional multicast routing technologies, multi-homing PEs attached to the same source can never create IP Multicast packet duplication if the PEs use a multi-homed Ethernet Segment (ES). Figure 3 illustrates this by showing two multi-homing PEs (PE1 and PE2) that are attached to the same source (S1). We assume that S1 is connected to an all-active ES by a layer-2 switch (SW1) with a Link Aggregation Group (LAG) to PE1 and PE2.

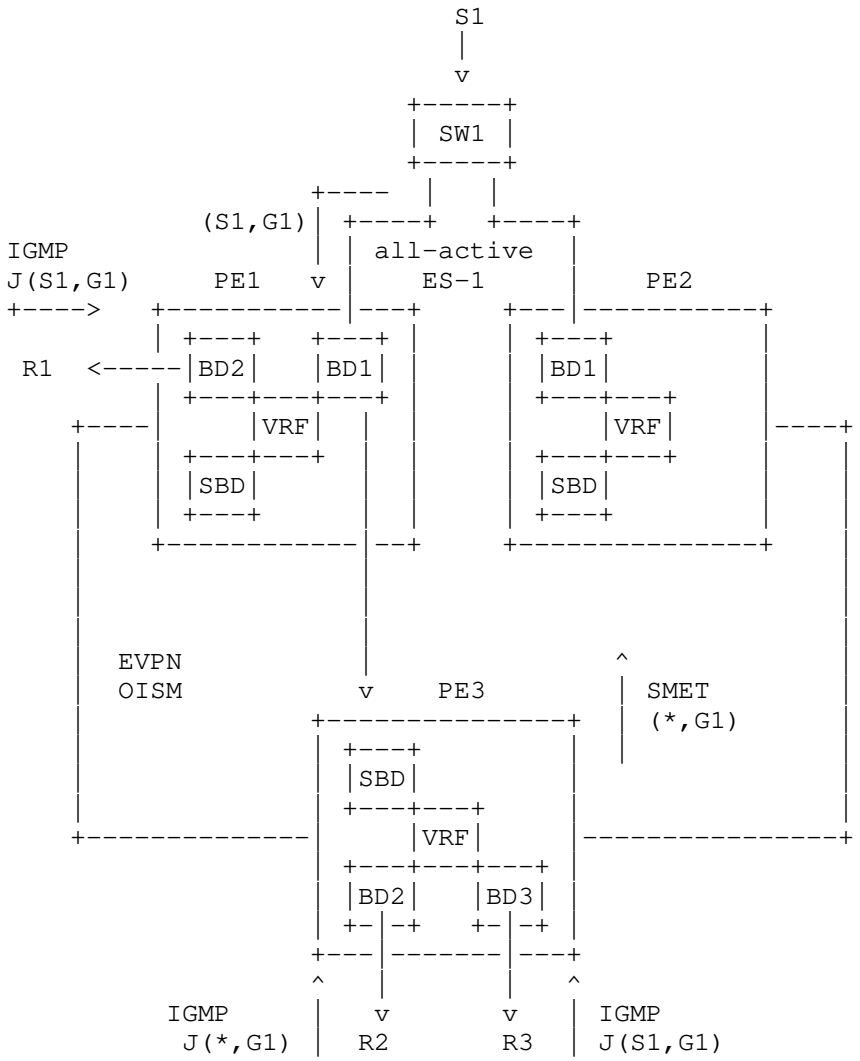

Figure 3: All-active Multi-homing and OISM

When receiving the (S1,G1) flow from S1, SW1 will choose only one link to send the flow, as per [RFC7432]. Assuming PE1 is the receiving PE on BD1, the IP Multicast flow will be forwarded as soon as BD1 creates multicast state for (S1,G1) or (*,G1). In the example of Figure 3, receivers R1, R2 and R3 are interested in the multicast flow to G1. R1 will receive (S1,G1) directly via the IRB interface as per [I-D.ietf-bess-evpn-irb-mcast]. Upon receiving IGMP reports from R2 and R3, PE3 will issue an SMET (*,G1) route that will create state in PE1's BD1. PE1 will therefore forward the IP Multicast flow

to PE3's SBD and PE3 will forward to R2 and R3, as per [I-D.ietf-bess-evpn-irb-mcast] procedures.

When IP Multicast source multi-homing is required, EVPN multi-homed Ethernet Segments MUST be used. EVPN multi-homing guarantees that only one Upstream PE will forward a given multicast flow at the time, avoiding packet duplication at the Downstream PEs. In addition, the SMET route for a given flow creates state in all the multi-homing Upstream PEs. Therefore, in case of failure on the Upstream PE forwarding the flow, the backup Upstream PE can forward the flow immediately.

This document assumes that multi-homing PEs attached to the same source always use multi-homed Ethernet Segments.

1.4. The Need for Redundant IP Multicast Sources in EVPN

While multi-homing PEs to the same IP Multicast G-source provides certain level of resiliency, multicast applications are often critical in the Operator's network and greater level of redundancy is required. This document assumes that:

a. Redundant G-sources for an SFG may exist in the EVPN tenant network. A Redundant G-source is a host or a router that sends an SFG in a tenant network where there is another host or router sending traffic to the same SFG.

b. Those redundant G-sources may be in the same BD or different BDs of the tenant. There must not be restrictions imposed on the location of the receiver systems either.

c. The redundant G-sources can be single-homed to only one EVPN PE or multi-homed to multiple EVPN PEs.

d. The EVPN PEs must avoid duplication of the same SFG on the receiver systems.

2. Solution Overview

An SFG is represented as (*,G) if any source that issues multicast traffic to G is a redundant G-source. Alternatively, this document allows an SFG to be represented as (S,G), where S is a prefix of any length. In this case, a source is considered a redundant G-source for the SFG if it is contained in the prefix. This document allows variable length prefixes in the Sources advertised in S-PMSI A-D routes only for the particular application of redundant G-sources.

There are two redundant G-source solutions described in this document:

- Warm Standby (WS) Solution
- Hot Standby (HS) Solution

The WS solution is considered an upstream-PE-based solution (since downstream PEs do not participate in the procedures), in which all the upstream PEs attached to redundant G-sources for an SFG represented by (*,G) or (S,G) will elect a "Single Forwarder" (SF) among themselves. Once a SF is elected, the upstream PEs add an Reverse Path Forwarding (RPF) check to the (*,G) or (S,G) state for the SFG:

- A non-SF upstream PE discards any (*,G)/(S,G) packets received over a local AC.
- The SF accepts and forwards any (*,G)/(S,G) packets it receives over a single local AC (for the SFG). In case (*,G)/(S,G) packets for the SFG are received over multiple local ACs, they will be discarded in all the local ACs but one. The procedure to choose the local AC that accepts packets is a local implementation matter.

A failure on the SF will result in the election of a new SF. The Election requires BGP extensions on the existing EVPN routes. These extensions and associated procedures are described in Section 3 and Section 4 respectively.

In the HS solution the downstream PEs are the ones avoiding the SFG duplication. The upstream PEs are aware of the locally attached G-sources and add a unique Ethernet Segment Identifier label (ESI-label) per SFG to the SFG packets forwarded to downstream PEs. The downstream PEs pull the SFG from all the upstream PEs attached to the redundant G-sources and avoid duplication on the receiver systems by adding an RPF check to the (*,G) state for the SFG:

- A downstream PE discards any (*,G) packets it receives from the "wrong G-source".
- The wrong G-source is identified in the data path by an ESI-label that is different than the ESI-label used for the selected G-source.
- Note that the ESI-label is used here for "ingress filtering" (at the egress/downstream PE) as opposed to the [RFC7432] "egress filtering" (at the egress/downstream PE) used in the split-horizon

procedures. In [RFC7432] the ESI-label indicates what egress ACs must be skipped when forwarding BUM traffic to the egress. In this document, the ESI-label indicates what ingress traffic must be discarded at the downstream PE.

The use of ESI-labels for SFGs forwarded by upstream PEs require some control plane and data plane extensions in the procedures used by [RFC7432] for multi-homing. Upon failure of the selected G-source, the downstream PE will switch over to a different selected G-source, and will therefore change the RPF check for the (*,G) state. The extensions and associated procedures are described in Section 3 and Section 5 respectively.

An operator should use the HS solution if they require a fast fail-over time and the additional bandwidth consumption is acceptable (SFG packets are received multiple times on the downstream PEs). Otherwise the operator should use the WS solution, at the expense of a slower fail-over time in case of a G-source or upstream PE failure. Besides bandwidth efficiency, another advantage of the WS solution is that only the upstream PEs attached to the redundant G-sources for the same SFG need to be upgraded to support the new procedures.

This document does not impose the support of both solutions on a system. If one solution is supported, the support of the other solution is OPTIONAL.

3. BGP EVPN Extensions

This document makes use of the following BGP EVPN extensions:

1. SFG flag in the Multicast Flags Extended Community

   The Single Flow Group (SFG) flag is a new bit requested to IANA out of the registry Multicast Flags Extended Community Flag Values. This new flag is set for S-PMSI A-D routes that carry a (*,G)/(S,G) SFG in the NLRI.

2. ESI Label Extended Community is used in S-PMSI A-D routes

   The HS solution requires the advertisement of one or more ESI Label Extended Communities [RFC7432] that encode the Ethernet Segment Identifier(s) associated to an S-PMSI A-D (*,G)/(S,G) route that advertises the presence of an SFG. Only the ESI Label value in the extended community is relevant to the procedures in this document. The Flags field in the extended community will be advertised as 0x00 and ignored on reception. [RFC7432] specifies that the ESI Label Extended Community is advertised along with the A-D per ES route. This documents extends the use of this

extended community so that it can be advertised multiple times (with different ESI values) along with the S-PMSI A-D route.

4. Warm Standby (WS) Solution for Redundant G-Sources

The general procedure is described as follows:

1. Configuration of the upstream PEs

   Upstream PEs (possibly attached to redundant G-sources) need to be configured to know which groups are carrying only flows from redundant G-sources, that is, the SFGs in the tenant domain. They will also be configured to know which local BDs may be attached to a redundant G-source. The SFGs can be configured for any source, E.g., SFG for "*", or for a prefix that contains multiple sources that will issue the same SFG, i.e., "10.0.0.0/30". In the latter case sources 10.0.0.1 and 10.0.0.2 are considered as Redundant G-sources, whereas 10.0.0.10 is not considered a redundant G-source for the same SFG.

   As an example:

   * PE1 is configured to know that G1 is an SFG for any source and redundant G-sources for G1 may be attached to BD1 or BD2.

   * Or PE1 can also be configured to know that G1 is an SFG for the sources contained in 10.0.0.0/30, and those redundant G-sources may be attached to BD1 or BD2.

2. Signaling the location of a G-source for a given SFG

   Upon receiving G-traffic for a configured SFG on a BD, an upstream PE configured to follow this procedure, e.g., PE1:

   * Originates an S-PMSI A-D (*,G)/(S,G) route for the SFG. An (*,G) route is advertised if the SFG is configured for any source, and an (S,G) route is advertised (where the Source can have any length) if the SFG is configured for a prefix.

   * The S-PMSI A-D route is imported by all the PEs attached to the tenant domain. In order to do that, the route will use the SBD-RT (Supplementary Broadcast Domain Route-Target) in addition to the BD-RT of the BD over which the G-traffic is received. The route SHOULD also carry a DF Election Extended Community (EC) and a flag indicating that it conveys an SFG. The DF Election EC and its use is specified in [RFC8584].

* The above S-PMSI A-D route MAY be advertised with or without PMSI Tunnel Attribute (PTA):

  + With no PTA if an I-PMSI or S-PMSI A-D with IR/AR/BIER are to be used.

  + With PTA in any other case.

* The S-PMSI A-D route is triggered by the first packet of the SFG and withdrawn when the flow is not received anymore. Detecting when the G-source is no longer active is a local implementation matter. The use of a timer is RECOMMENDED. The timer is started when the traffic to G1 is not received. Upon expiration of the timer, the PE will withdraw the route

3. Single Forwarder (SF) Election

   If the PE with a local G-source receives one or more S-PMSI A-D routes for the same SFG from a remote PE, it will run a Single Forwarder (SF) Election based on the information encoded in the DF Election EC. Two S-PMSI A-D routes are considered for the same SFG if they are advertised for the same tenant, and their Multicast Source Length, Multicast Source, Multicast Group Length and Multicast Group fields match.

   1. A given DF Alg can only be used if all the PEs running the DF Alg have consistent input. For example, in an OISM network, if the redundant G-sources for an SFG are attached to BDs with different Ethernet Tags, the Default DF Election Alg MUST NOT be used.

   2. In case the there is a mismatch in the DF Election Alg or capabilities advertised by two PEs competing for the SF, the lowest PE IP address (given by the Originator Address in the S- PMSI A-D route) will be used as a tie-breaker.

4. RPF check on the PEs attached to a redundant G-source

   All the PEs with a local G-source for the SFG will add an RPF check to the (*,G)/(S,G) state for the SFG. That RPF check depends on the SF Election result:

   1. The non-SF PEs discard any (*,G)/(S,G) packets for the SFG received over a local AC.

   2. The SF accepts any (*,G)/(S,G) packets for the SFG it receives over one (and only one) local AC.

The solution above provides redundancy for SFGs and it does not require an upgrade of the downstream PEs (PEs where there is certainty that no redundant G-sources are connected). Other G-sources for non-SFGs may exist in the same tenant domain. This document does not change the existing procedures for non-SFG G-sources.

The redundant G-sources can be single-homed or multi-homed to a BD in the tenant domain. Multi-homing does not change the above procedures.

Section 4.1 and Section 4.2 show two examples of the WS solution.

4.1. WS Example in an OISM Network

Figure 4 illustrates an example in which S1 and S2 are redundant G-sources for the SFG (*,G1).

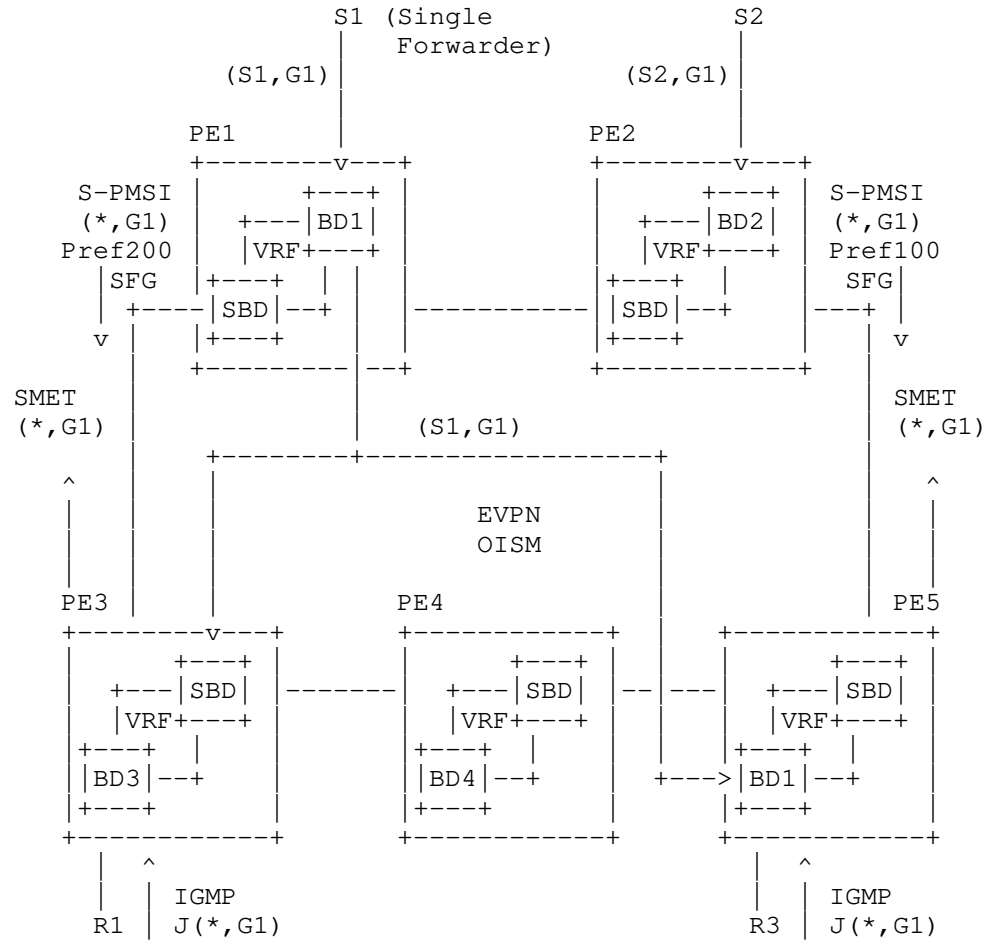

Figure 4: WS Solution for Redundant G-Sources

The WS solution works as follows:

1. Configuration of the upstream PEs, PE1 and PE2

   PE1 and PE2 are configured to know that G1 is an SFG for any source and redundant G-sources for G1 may be attached to BD1 or BD2, respectively.

2. Signaling the location of S1 and S2 for (*,G1)

   Upon receiving (S1,G1) traffic on a local AC, PE1 and PE2 originate S-PMSI A-D (*,G1) routes with the SBD-RT, DF Election

Extended Community (EC) and a flag indicating that it conveys an SFG.

3. Single Forwarder (SF) Election

   Based on the DF Election EC content, PE1 and PE2 elect an SF for (*,G1). Assuming both PEs agree on e.g., Preference based Election as the algorithm to use [DF-PREF], and PE1 has a higher preference, PE1 becomes the SF for (*,G1).

4. RPF check on the PEs attached to a redundant G-source

   A. The non-SF, PE2, discards any (*,G1) packets received over a local AC.

   B. The SF, PE1 accepts (*,G1) packets it receives over one (and only one) local AC.

The end result is that, upon receiving reports for (*,G1) or (S,G1), the downstream PEs (PE3 and PE5) will issue SMET routes and will pull the multicast SFG from PE1, and PE1 only. Upon a failure on S1, the AC connected to S1 or PE1 itself will trigger the S-PMSI A-D (*,G1) withdrawal from PE1 and PE2 will be promoted to SF.

4.2. WS Example in a Single-BD Tenant Network

Figure 5 illustrates an example in which S1 and S2 are redundant G-sources for the SFG (*,G1), however, now all the G-sources and receivers are connected to the same BD1 and there is no SBD.

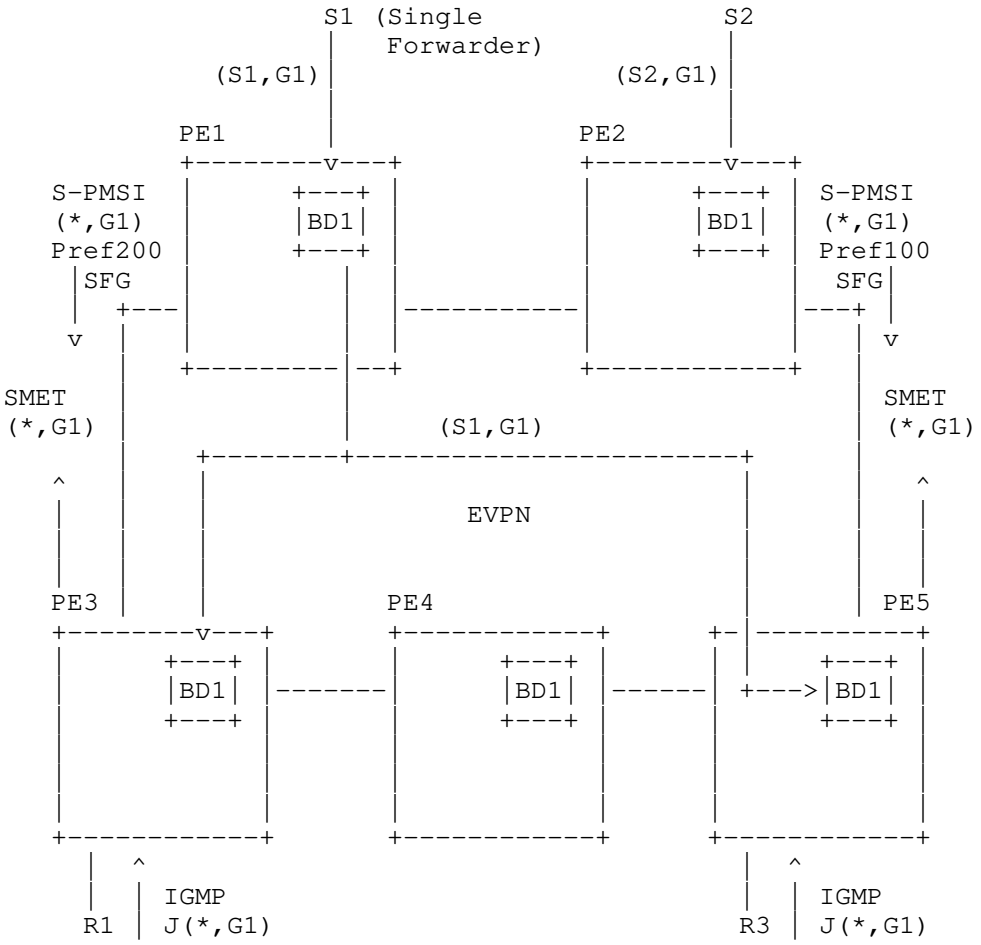

Figure 5: WS Solution for Redundant G-Sources in the same BD

The same procedure as in Section 4.1 is valid here, being this a sub-case of the one in Section 4.1. Upon receiving traffic for the SFG G1, PE1 and PE2 advertise the S-PMSI A-D routes with BD1-RT only, since there is no SBD.

5. Hot Standby (HS) Solution for Redundant G-Sources

If fast-failover is required upon the failure of a G-source or PE attached to the G-source and the extra bandwidth consumption in the tenant network is not an issue, the HS solution should be used. The procedure is as follows:

1. Configuration of the PEs

As in the WS case, the upstream PEs where redundant G-sources may exist need to be configured to know which groups (for any source or a prefix containing the intended sources) are carrying only flows from redundant G-sources, that is, the SFGs in the tenant domain.

In addition (and this is not done in WS mode), the individual redundant G-sources for an SFG need to be associated with an Ethernet Segment (ES) on the upstream PEs. This is irrespective of the redundant G-source being multi-homed or single-homed. Even for single-homed redundant G-sources the HS procedure relies on the ESI labels for the RPF check on downstream PEs. The term "S-ESI" is used in this document to refer to an ESI associated to a redundant G-source.

Contrary to what is specified in the WS method (that is transparent to the downstream PEs), the support of the HS procedure is required not only on the upstream PEs but also on all downstream PEs connected to the receivers in the tenant network. The downstream PEs do not need to be configured to know the connected SFGs or their ESIs, since they get that information from the upstream PEs. The downstream PEs will locally select an ESI for a given SFG, and will program an RPF check to the (*,G)/(S,G) state for the SFG that will discard (*,G)/(S,G) packets from the rest of the ESIs. The selection of the ESI for the SFG is based on local policy.

2. Signaling the location of a G-source for a given SFG and its association to the local ESIs

   Based on the configuration in step 1, an upstream PE configured to follow the HS procedures:

   A. Advertises an S-PMSI A-D (*,G)/(S,G) route per each configured SFG. These routes need to be imported by all the PEs of the tenant domain, therefore they will carry the BD-RT and SBD-RT (if the SBD exists). The route also carries the ESI Label Extended Communities needed to convey all the S-ESIs associated to the SFG in the PE.

   B. The S-PMSI A-D route will convey a PTA in the same cases as in the WS procedure.

   C. The S-PMSI A-D (*,G)/(S,G) route is triggered by the configuration of the SFG and not by the reception of G-traffic.

3. Distribution of DCB (Domain-wide Common Block) ESI-labels and G-source ES routes

   An upstream PE advertises the corresponding ES, A-D per EVI and A-D per ES routes for the local S-ESIs.

   A. ES routes are used for regular DF Election for the S-ES. This document does not introduce any change in the procedures related to the ES routes.

   B. The A-D per EVI and A-D per ES routes MUST include the SBD-RT since they have to be imported by all the PEs in the tenant domain.

   C. The A-D per ES routes convey the S-ESI labels that the downstream PEs use to add the RPF check for the (*,G)/(S,G) associated to the SFGs. This RPF check requires that all the packets for a given G-source are received with the same S-ESI label value on the downstream PEs. For example, if two redundant G-sources are multi-homed to PE1 and PE2 via S-ES-1 and S-ES-2, PE1 and PE2 MUST allocate the same ESI label "Lx" for S-ES-1 and they MUST allocate the same ESI label "Ly" for S-ES-2. In addition, Lx and Ly MUST be different. These ESI labels are Domain-wide Common Block (DCB) labels and follow the allocation procedures in [I-D.zzhang-bess-mvpn-evpn-aggregation-label].

4. Processing of A-D per ES/EVI routes and RPF check on the downstream PEs

   The A-D per ES/EVI routes are received and imported in all the PEs in the tenant domain. The processing of the A-D per ES/EVI routes on a given PE depends on its configuration:

   A. The PEs attached to the same BD of the BD-RT that is included in the A-D per ES/EVI routes will process the routes as in [RFC7432] and [RFC8584]. If the receiving PE is attached to the same ES as indicated in the route, [RFC7432] split-horizon procedures will be followed and the DF Election candidate list may be modified as in [RFC8584] if the ES supports the AC-DF capability.

   B. The PEs that are not attached to the BD-RT but are attached to the SBD of the received SBD-RT, will import the A-D per ES/EVI routes and use them for redundant G-source mass withdrawal, as explained later.

C. Upon importing A-D per ES routes corresponding to different S-ESes, a PE MUST select a primary S-ES and add an RPF check to the (*,G)/(S,G) state in the BD or SBD. This RPF check will discard all ingress packets to (*,G)/(S,G) that are not received with the ESI-label of the primary S-ES. The selection of the primary S-ES is a matter of local policy.

5. G-traffic forwarding for redundant G-sources and fault detection

Assuming there is (*,G) or (S,G) state for the SFG with OIF (Ouput Interface) list entries associated to remote EVPN PEs, upon receiving G-traffic on a S-ES, the upstream PE will add a S-ESI label at the bottom of the stack before forwarding the traffic to the remote EVPN PEs. This label is allocated from a DCB as described in step 3. If P2MP or BIER PMSIs are used, this is not adding any new data path procedures on the upstream PEs (except that the ESI-label is allocated from a DCB). However, if IR/AR are used, this document extends the [RFC7432] procedures by pushing the S-ESI labels not only on packets sent to the PEs that shared the ES but also to the rest of the PEs in the tenant domain. This allows the downstream PEs to receive all the multicast packets from the redundant G-sources with a S-ESI label (irrespective of the PMSI type and the local ESes), and discard any packet that conveys a S-ESI label different from the primary S-ESI label (that is, the label associated to the selected primary S-ES), as discussed in step 4.

If the last A-D per EVI or the last A-D per ES route for the primary S-ES is withdrawn, the downstream PE will immediately select a new primary S-ES and will change the RPF check. Note that if the S-ES is re-used for multiple tenant domains by the upstream PEs, the withdrawal of all the A-D per-ES routes for a S-ES provides a mass withdrawal capability that makes a downstream PE to change the RPF check in all the tenant domains using the same S-ES.

The withdrawal of the last S-PMSI A-D route for a given (*,G)/(S,G) that represents a SFG SHOULD make the downstream PE remove the S-ESI label based RPF check on (*,G)/(S,G).

5.1. Use of BFD in the HS Solution

In addition to using the state of the A-D per EVI, A-D per ES or S-PMSI A-D routes to modify the RPF check on (*,G)/(S,G) as discussed in Section 5, Bidirectional Forwarding Detection (BFD) protocol MAY be used to find the status of the multipoint tunnels used to forward the SFG from the redundant G-sources.

The BGP-BFD Attribute is advertised along with the S-PMSI A-D or IMET routes (depending on whether I-PMSI or S-PMSI trees are used) and the procedures described in [EVPN-BFD] are used to bootstrap multipoint BFD sessions on the downstream PEs.

5.2. HS Example in an OISM Network

Figure 6 illustrates the HS model in an OISM network. Consider S1 and S2 are redundant G-sources for the SFG (*,G1) in BD1 (any source using G1 is assumed to transmit an SFG). S1 and S2 are (all-active) multi-homed to upstream PEs, PE1 and PE2. The receivers are attached to downstream PEs, PE3 and PE5, in BD3 and BD1, respectively. S1 and S2 are assumed to be connected by a LAG to the multi-homing PEs, and the multicast traffic can use the link to either upstream PE. The diagram illustrates how S1 sends the G-traffic to PE1 and PE1 forwards to the remote interested downstream PEs, whereas S2 sends to PE2 and PE2 forwards further. In this HS model, the interested downstream PEs will get duplicate G-traffic from the two G-sources for the same SFG. While the diagram shows that the two flows are forwarded by different upstream PEs, the all-active multi-homing procedures may cause that the two flows come from the same upstream PE. Therefore, finding out the upstream PE for the flow is not enough for the downstream PEs to program the required RPF check to avoid duplicate packets on the receiver.

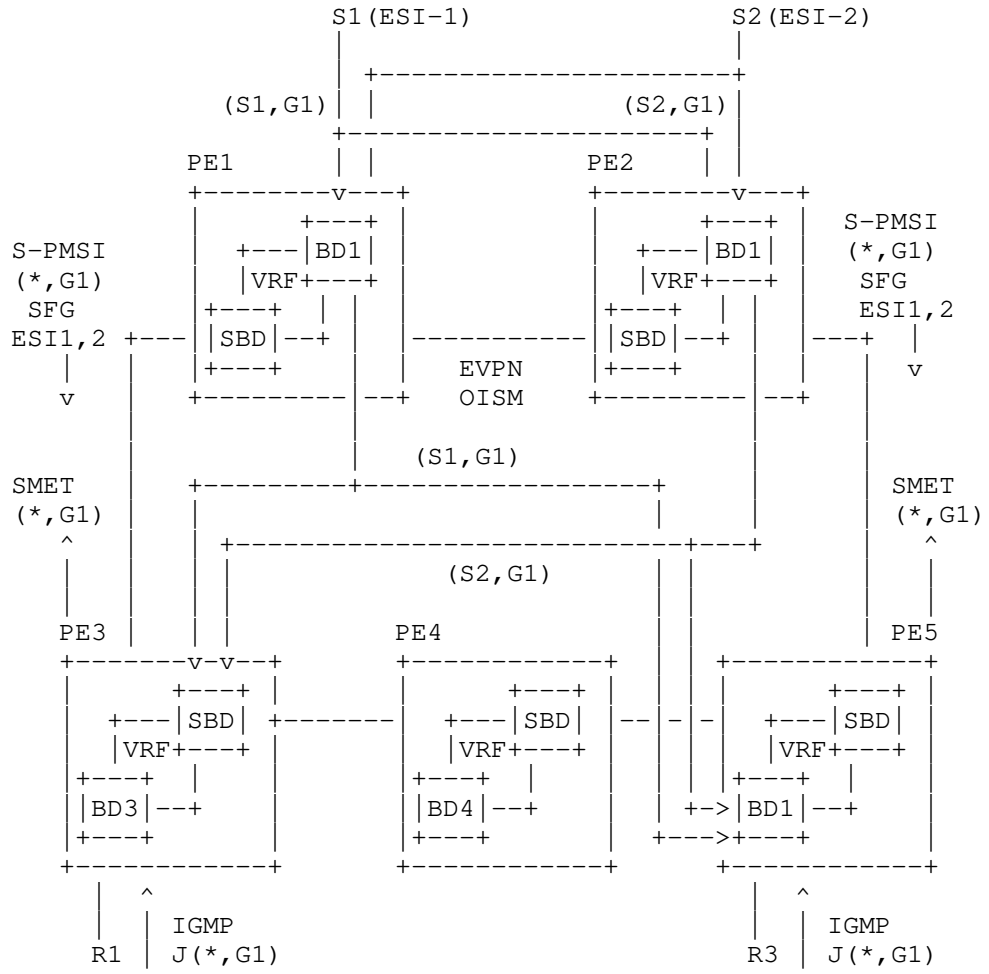

Figure 6: HS Solution for Multi-homed Redundant G-Sources in OISM

In this scenario, the HS solution works as follows:

1. Configuration of the upstream PEs, PE1 and PE2

   PE1 and PE2 are configured to know that G1 is an SFG for any source (a source prefix length could have been configured instead) and the redundant G-sources for G1 use S-ESIs ESI-1 and ESI-2 respectively. Both ESes are configured in both PEs and the ESI value can be configured or auto-derived. The ESI-label values are allocated from a DCB [I-D.zzhang-bess-mvpn-evpn-aggregation-label] and are configured

either locally or by a centralized controller. We assume ESI-1 is configured to use ESI-label-1 and ESI-2 to use ESI-label-2.

The downstream PEs, PE3, PE4 and PE5 are configured to support HS mode and select the G-source with e.g., lowest ESI value.

2. PE1 and PE2 advertise S-PMSI A-D (*,G1) and ES/A-D per ES/EVI routes

   Based on the configuration of step 1, PE1 and PE2 advertise an S-PMSI A-D (*,G1) route each. The route from each of the two PEs will include TWO ESI Label Extended Communities with ESI-1 and ESI-2 respectively, as well as BD1-RT plus SBD-RT and a flag that indicates that (*,G1) is an SFG.

   In addition, PE1 and PE2 advertise ES and A-D per ES/EVI routes for ESI-1 and ESI-2. The A-D per ES and per EVI routes will include the SBD-RT so that they can be imported by the downstream PEs that are not attached to BD1, e.g., PE3 and PE4. The A-D per ES routes will convey ESI-label-1 for ESI-1 (on both PEs) and ESI-label-2 for ESI-2 (also on both PEs).

3. Processing of A-D per ES/EVI routes and RPF check

   PE1 and PE2 received each other’s ES and A-D per ES/EVI routes. Regular [RFC7432] [RFC8584] procedures will be followed for DF Election and programming of the ESI-labels for egress split-horizon filtering. PE3/PE4 import the A-D per ES/EVI routes in the SBD. Since PE3 has created a (*,G1) state based on local interest, PE3 will add an RPF check to (*,G1) so that packets coming with ESI-label-2 are discarded (lowest ESI value is assumed to give the primary S-ES).

4. G-traffic forwarding and fault detection

   PE1 receives G-traffic (S1,G1) on ES-1 that is forwarded within the context of BD1. Irrespective of the tunnel type, PE1 pushes ESI-label-1 at the bottom of the stack and the traffic gets to PE3 and PE5 with the mentioned ESI-label (PE4 has no local interested receivers). The G-traffic with ESI-label-1 passes the RPF check and it is forwarded to R1. In the same way, PE2 sends (S2,G1) with ESI-label-2, but this G-traffic does not pass the RPF check and gets discarded at PE3/PE5.

   If the link from S1 to PE1 fails, S1 will forward the (S1,G1) traffic to PE2 instead. PE1 withdraws the ES and A-D routes for ESI-1. Now both flows will be originated by PE2, however the RPF checks don’t change in PE3/PE5.

If subsequently, the link from S1 to PE2 fails, PE2 also withdraws the ES and A-D routes for ESI-1. Since PE3 and PE5 have no longer A-D per ES/EVI routes for ESI-1, they immediately change the RPF check so that packets with ESI-label-2 are now accepted.

Figure 7 illustrates a scenario where S1 and S2 are single-homed to PE1 and PE2 respectively. This scenario is a sub-case of the one in Figure 6. Now ES-1 only exists in PE1, hence only PE1 advertises the A-D per ES/EVI routes for ESI-1. Similarly, ES-2 only exists in PE2 and PE2 is the only PE advertising A-D routes for ESI-2. The same procedures as in Figure 6 applies to this use-case.

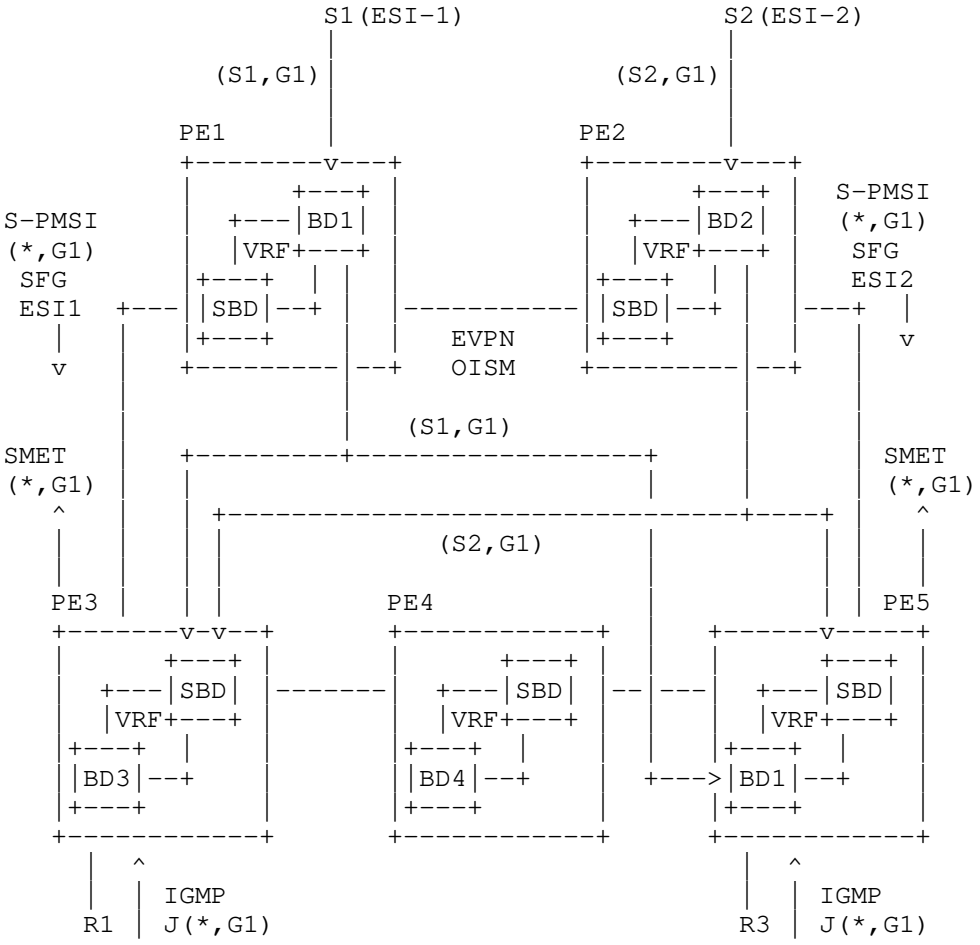

Figure 7: HS Solution for single-homed Redundant G-Sources in OISM

5.3. HS Example in a Single-BD Tenant Network

Irrespective of the redundant G-sources being multi-homed or single-homed, if the tenant network has only one BD, e.g., BD1, the procedures of Section 5.2 still apply, only that routes do not include any SBD-RT and all the procedures apply to BD1 only.

6. Security Considerations

The same Security Considerations described in [I-D.ietf-bess-evpn-irb-mcast] are valid for this document.

From a security perspective, out of the two methods described in this document, the WS method is considered lighter in terms of control plane and therefore its impact is low on the processing capabilities of the PEs. The HS method adds more burden on the control plane of all the PEs of the tenant with sources and receivers.

7. IANA Considerations

IANA is requested to allocate a Bit in the Multicast Flags Extended Community to indicate that a given (*,G) or (S,G) in an S-PMSI A-D route is associated with an SFG.

8. References

8.1. Normative References

[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>.

[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>.

[I-D.ietf-bess-evpn-igmp-mld-proxy] Sajassi, A., Thoria, S., Patel, K., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-05 (work in progress), April 2020.

[I-D.ietf-bess-evpn-irb-mcast]
Lin, W., Zhang, Z., Drake, J., Rosen, E., Rabadan, J., and A. Sajassi, "EVPN Optimized Inter-Subnet Multicast (OISM) Forwarding", draft-ietf-bess-evpn-irb-mcast-05 (work in progress), October 2020.

[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>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

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

[I-D.zzhang-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", draft-zzhang-bess-mvpn-evpn-aggregation-label-01 (work in progress), April 2018.

8.2. Informative References

[EVPN-RT5]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", internet-draft ietf-bess-evpn-prefix-advertisement-11.txt, May 2018.

[EVPN-BUM]
Zhang, Z., Lin, W., Rabadan, J., and K. Patel, "Updates on EVPN BUM Procedures", internet-draft ietf-bess-evpn-bum-procedure-updates-06, June 2019.

[DF-PREF] Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., and S. Mohanty, "Preference-based EVPN DF Election", internet-draft ietf-bess-evpn-pref-df-04.txt, June 2019.

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

[EVPN-BFD]
Govindan, V., Mallik, M., Sajassi, A., and G. Mirsky, "Fault Management for EVPN networks", internet-draft ietf-bess-evpn-bfd-01.txt, October 2020.

Appendix A. Acknowledgments

The authors would like to thank Mankamana Mishra and Ali Sajassi for their review and valuable comments.

Appendix B. Contributors

Authors' Addresses

Jorge Rabadan (editor)
Nokia
777 Middlefield Road
Mountain View, CA 94043
USA

Email: jorge.rabadan@nokia.com

Jayant Kotalwar
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA

Email: jayant.kotalwar@nokia.com

Senthil Sathappan
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA

Email: senthil.sathappan@nokia.com

Zhaohui Zhang
Juniper Networks

Email: zzhang@juniper.net

Wen Lin
Juniper Networks

Email: wlin@juniper.net

Eric C. Rosen
Individual

Email: erosen52@gmail.com