BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: 13 April 2024

J. Rabadan, Ed.
J. Kotalwar
S. Sathappan
Nokia
Z. Zhaohui
Juniper Networks
A. Sajassi
M. Mishra
Cisco Systems
11 October 2023

PIM Proxy in EVPN Networks
draft-skr-bess-evpn-pim-proxy-02

Abstract

Ethernet Virtual Private Networks are becoming prevalent in Data Centers, Data Center Interconnect (DCI) and Service Provider VPN applications. One of the goals that EVPN pursues is the reduction of flooding and the efficiency of CE-based control plane procedures in Broadcast Domains. Examples of this are Proxy ARP/ND and IGMP/MLD Proxy. This document complements the latter, describing the procedures required to minimize the flooding of PIM messages in EVPN Broadcast Domains, and optimize the IP Multicast delivery between PIM routers.

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 13 April 2024.

Copyright Notice

Table of Contents

1. Introduction

Ethernet Virtual Private Networks [RFC7432] are becoming prevalent in Data Centers, Data Center Interconnect (DCI) and Service Provider VPN applications. One of the goals that EVPN pursues is the reduction of flooding and the efficiency of CE-based control plane procedures in Broadcast Domains. Examples of this are [RFC9161] for improving the efficiency of CE's ARP/ND protocols, and [RFC9251] for IGMP/MLD

protocols.

This document focuses on optimizing the behavior of PIM in EVPN Broadcast Domains and re-uses some procedures of [RFC9251]. The reader is also advised to check out [RFC8220] to understand certain aspects of the procedures of PIM Join/Prune messages received on Attachment Circuits (ACs).

Section 4 describes the PIM Proxy procedures that the implementation should follow, including:

* The use of EVPN to suppress the flooding of PIM Hello messages in shared Broadcast Domains. The benefit of this is twofold:
  - PIM Hello messages will ONLY be flooded to Attachment Circuits that are connected to PIM routers, as opposed to all the CEs and hosts in the Broadcast Domain.
  - Soft-state PIM Hello messages will be replaced by hard-state BGP messages that don't need to be refreshed periodically.
* The use of EVPN to discover IGMP Queriers, while avoiding the flooding of IGMP Queries in the core.
* The procedures to proxy PIM Join/Prune messages and replace them by hard-state EVPN routes that don't need to be refreshed periodically. By using BGP EVPN to propagate both, Hello and Join/Prune messages, we also avoid out-of-order delivery between both types of PIM messages.
* This document also describes an EVPN based procedure so that the PIM routers connected to the shared Broadcast Domain don't need to run any PIM Assert procedure. PIM Assert procedures may be expensive for PIM routers in terms of resource consumption. With this procedure, there is no PIM Assert needed on PIM routers.
* The use of procedures similar to the ones defined in [EVPN-IGMP-MLD-PROXY] to synchronize multicast states among the PEs in the same Ethernet Segment.

Section 5 describes the interaction of PIM Proxy with IGMP Proxy PEs and Multicast Sources connected to the same EVPN Broadcast Domain.

Section 6 defines the BGP Information Model that this document requires to address the PIM Proxy procedures.

This document assumes the reader is familiar with PIM and IGMP protocols.

2. Conventions used in this document

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.

3. Terminology

This section summarizes the terminology that is used throughout the rest of the document.

* AC: Attachment Circuit or logical interface associated to a given Broadcast Domain. To determine the AC on which a packet arrived, the PE examines the combination of a physical port and VLAN tags (where the VLAN tags can be individual c-tags, s-tags or ranges of both).

* EVI: EVPN Instance.

* EVPN Broadcast Domain: it refers to an EVI in case of VLAN-based and VLAN-bundle interfaces. It refers to a Bridge Domain identified by an Ethernet-Tag (in the control plane) in case of VLAN-Aware Bundle interfaces.

* PIM-DM: Protocol Independent Multicast - Dense Mode.

* PIM-SM: Protocol Independent Multicast - Sparse Mode.

* PIM-SSM: Protocol Independent Multicast - Source Specific Mode.

* S: IP address of the multicast source.

* G: IP address of the multicast group.

* N: Upstream neighbor field in a Join/Prune/Graft message.

* PIM J/P: PIM Join/Prune messages.

* RP: PIM Rendezvous Point.

* MRD route: Multicast Router Discovery.

* PIM Nbr: PIM Neighbor.

4. PIM Proxy Operation in EVPN Broadcast Domains

This section describes the operation of PIM Proxy in EVPN Broadcast Domains (BDs). Figure 1 depicts an EVPN Broadcast Domain defined in four PEs that are connected to PIM routers. This example will be used throughout this section and assumes both R4 and R5 are PIM Upstream Neighbors for PIM routers R1, R2 and R3 and multicast group G1. In this situation, the PIM multicast traffic flows from R4 or R5 to R1, R2 and R3. The PIM Join/Prune signaling will flow in the opposite direction. From a terminology perspective, we consider PE1 and PE2 as egress or downstream PEs, whereas PE3 and PE4 are ingress or upstream PEs.

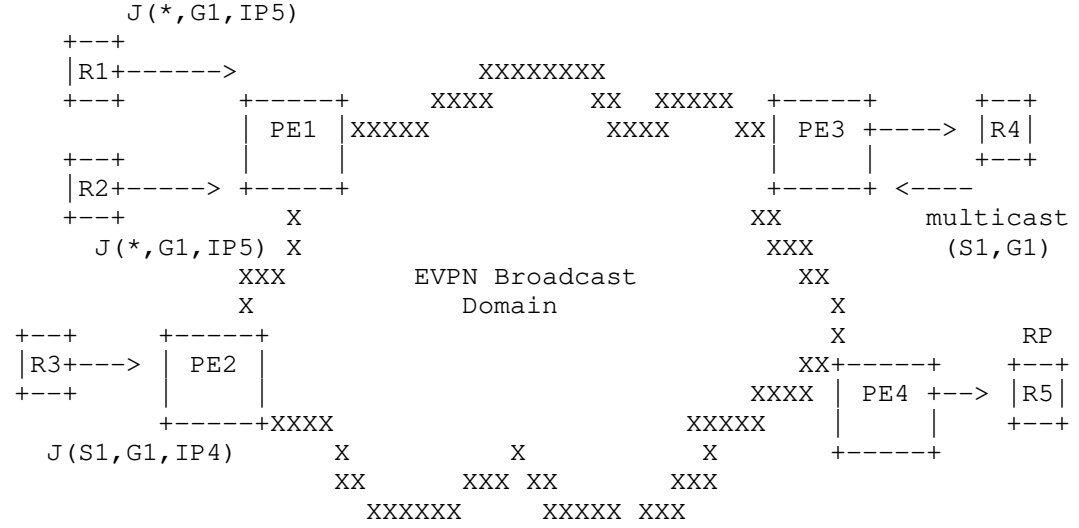

Figure 1: PIM Routers connected by an EVPN Broadcast Domain

It is important to note that any Router's PIM message not explicitly specified in this document will be forwarded by the PEs normally, in the data path, as a unicast or multicast packet.

4.1. Multicast Router Discovery Procedures in EVPN

The procedures defined in this section make use of the Multicast Router Discovery (MRD) route described in section 4 and are OPTIONAL. An EVPN router not implementing this specification will transparently flood PIM Hello messages and IGMP Queries to remote PEs.

4.1.1. Discovering PIM Routers

As described in [RFC7761] for shared LANs, an EVPN Broadcast Domain may have multiple PIM routers connected to it and a single one of these routers, the DR, will act on behalf of directly connected hosts with respect to the PIM-SM protocol. The DR election, as well as discovery and negotiation of options in PIM, is performed using Hello messages. PIM Hello messages are periodically exchanged and flooded in EVPN Broadcast Domains that don't follow this specification. When PIM Proxy is enabled, an EVPN PE will snoop PIM Hello messages and forward them only to local ACs where PIM routers have been detected. This document assumes that all the procedures defined in [RFC8220] to snoop PIM Hellos on local ACs and build the PIM Neighbor DB on the PEs are followed. PIM Hello messages MUST NOT be forwarded to remote EVPN PEs though.

Using Figure 1 as an example, the PIM Proxy operation for Hello messages is as follows:

1. The arrival of a new PIM Hello message at e.g. PE1 will trigger an MRD route advertisement including:

    * The IP address and length of the multicast router that issued the Hello message. E.g. R1's IP address and length.

    * The DR Priority copied from the Hello DR Priority TLV.

    * Q flag set (if the multicast router is a Querier).

    * P flag set that indicates the router is PIM capable.

2. All other PEs import the MRD route and do the following:

    * Add the multicast router address to the PIM Neighbor Database (PIM Nbr DB) associated to the Originator Router Address.

    * Generate a PIM hello where the IP Source Address is the Multicast Router IP and the DR Priority is copied from the route. This PIM hello is sent to all the local ACs connected to a PIM router. For example, PE3 will send the generated hello message to R4.

3. Each PE will build its PIM Nbr DB out of the local PIM hello messages and/or remote MRD routes. The PIM hello timers and other hello parameters are not propagated in the MRD routes.

* The timers are handled locally by the PE and as per [RFC7761]. This is valid for the hold_time (when a PIM router or PE receives a hello message, resets the neighbor-expiry timer), and other timers.

* The Generation ID option is also processed locally on the PE, as well as the Generation ID changes for a given multicast router. It is not propagated in the MRD route.

* Procedures described in [RFC7761] are used to remove a local AC PIM router from the PIM Nbr DB. When a local router is removed from the DB, the MRD route is withdrawn. If the local router is still sending Queries, the route is updated with flags P=0 and Q=1. Upon receiving the update, the other PEs will remove the router from the PIM Nbr DB but not from the list of queriers.

4. Based on regular PIM DR election procedures (highest DR Priority or highest IP), each PE is aware of who the DR is for the BD. For more information, refer to section "3. Interaction with IGMP- snooping and Sources".

4.1.2. Discovering IGMP Queriers

In (EVPN) Broadcast Domains that are shared among not only PIM routers but also IGMP hosts, one or more PIM routers will also be configured as IGMP Queriers. The proxy Querier mechanism described in [RFC9251] suppresses the flooding of queries on the Broadcast Domain, by using PE generated Queries from an anycast IP address.

While the proxy Querier mechanism works in most of the use-cases, sometimes it is desired to have a more transparent behavior and propagate existing multicast router IGMP Queries as opposed to "blindly" querying all the hosts from the PEs. The MRD route defined in Section 6 can be used for that purpose.

When the discovered local PIM router is also sending IGMP Queries, the PE will issue an MRD route for the multicast router with both Q (IGMP Querier) and P (PIM router) flags set. Note that the PE may set both flags or only one of them, depending on the capabilities of the local router.

A PE receiving an MRD route with Q=1 will generate IGMP Query messages, using the multicast router IP address encoded in the received MRD route. If more than one IGMP Queriers exist in the EVI, the PE receiving the MRD routes with Q=1 will select the lower IP address, as per [RFC2236]. Note that, upon receiving the MRD routes with Q=1, the PE must generate IGMP Queries and forward them to all

the local ACs. Other Queriers listening to these received Query messages will stop sending Queries if they are no longer the selected Querier, as per [RFC2236]. This procedure allows the EVPN PEs to act as proxy Queriers, but using the IP address of the best existing IGMP Querier in the EVPN Broadcast Domain. This can help IGMP hosts troubleshoot any issues on the IGMP routers and check their connectivity to them.

4.2. PIM Join/Prune Proxy Procedures

The procedures defined in this section make use of the Multicast Router Discovery (MRD) route described in section 4 and are OPTIONAL. An EVPN router not implementing this specification will transparently flood PIM Hello messages and IGMP Queries to remote PEs.

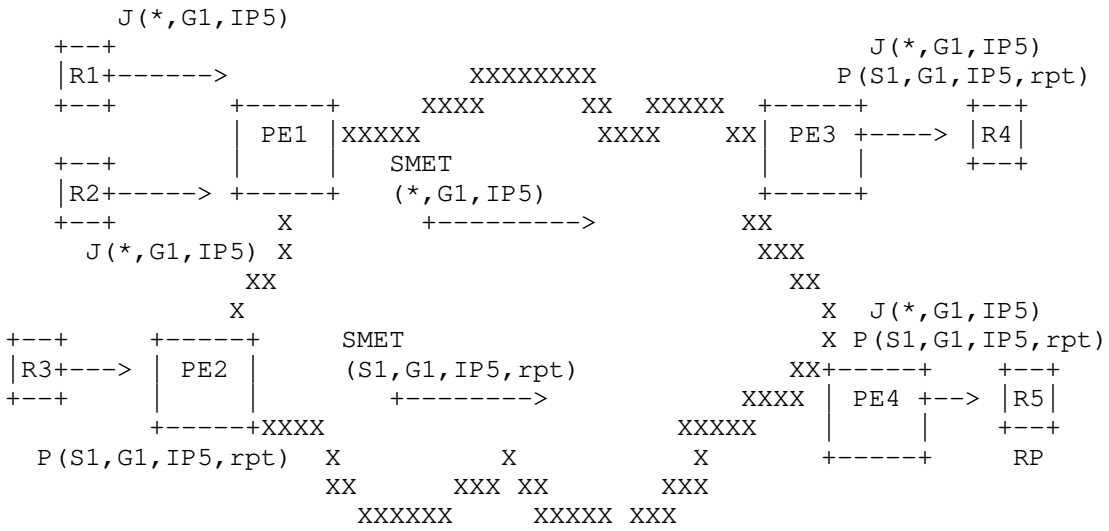

Figure 2: Proxy PIM Join/Prune in EVPN

PIM J/P messages are sent by the routers towards upstream sources and RPs:

* (*,G) is used in Join/Prune messages that are sent towards the RP for the specified group.

* (S,G) used in Join/Prune messages sent towards the specified source.

* (S,G,rpt) is used in Join/Prune messages sent towards the RP. We refer to this as RPT message and the Prune message always precedes the Join message. The typical sequence of PIM messages (for a group) seen in a BD connecting PIM routers is the following:

  a. (*,G) Join issued by a downstream router to the RP (to join the RP Tree).

  b. (S,G) Join issued by a downstream router switching to the SPT.

  c. (S,G,rpt) Prune issued by a downstream router to the RP to prune a specific source from the RPT.

  d. (S,G) Prune issued by a downstream router no longer interested in the SPT.

  e. (S,G,rpt) Join issued by a downstream router interested (again) in the RPT for (S,G).

The Proxy PIM procedures for Join/Prune messages are summarized as follows:

1. Downstream PE procedures:

   * A downstream PE will snoop PIM Join/Prune messages and won't forward them to remote PEs.

   * Triggered by the reception of the PIM Join message, a downstream PE will advertise an SMET route, including the source, group and Upstream Neighbor as received from the PIM Join message. A single SMET route is advertised per source, group, with the P flag set. As an example, in Figure 2, PE1 receives two PIM Join messages for the same source, group and Upstream Neighbor, however PE1 advertises a single SMET route.

   * When the last connected router sends a PIM Prune message for a given source, group and Upstream Neighbor and the state is removed, the PE will withdraw the SMET route (note that the state is removed once the prune-pend timer expires).

   * SMET routes must always be generated upon receiving a PIM Join message, irrespective of the location of the Upstream Neighbor and even if the Upstream Neighbor is local to the PE.

* A downstream PE receiving a PIM Prune (S,G,rpt) message will trigger an RPT-Prune route for the source and group. Subsequently, if the downstream PE receives a PIM Join (S,G,rpt) to cancel the previous Prune (S,G,rpt) and keep pulling the multicast traffic from the RPT, the downstream PE will withdraw the RPT-Prune route.

* PIM Timers are handled locally. If the holdtime expires for a local Join the PE withdraws the SMET route.

2. Upstream PE procedures:

* A received SMET route with P=1 will add state for the source and group and will generate a PIM Join message for the source, group that will be forwarded to all the local AC PIM routers.

* A received SMET route withdrawal will remove the state and generate a PIM Prune message for the source, group and upstream neighbor that will be forwarded to all the local AC PIM routers.

* A received RPT-Prune route for (S,G) will generate a PIM Prune (S,G,rpt) message that will be forwarded to all the local AC PIM routers.

* A received RPT-Prune withdrawal for (S,G) will generate a PIM Join (S,G,rpt) message that will be forwarded to all the local AC PIM routers.

It is important to note that, compared to a solution that does not snoop PIM messages and does not use BGP to propagate states in the core, this EVPN PIM Proxy solution will add some latency derived from the procedures described in this document.

4.3. PIM Assert Optimization

The PIM Assert process described in [RFC7761] is intense in terms of resource consumption in the PIM routers, however it is needed in case PIM routers share a multi-access transit LAN. The use of PIM Proxy for EVPN BDs can minimize and even suppress the need for PIM Assert as described in this section.

As a refresher, the PIM Assert procedures are needed to prevent two or more Upstream PIM routers from forwarding the same multicast content to the group of Downstream PIM routers sharing the same (EVPN) Broadcast Domain. This multicast packet duplication may happen in any of the following cases:

* Two or more Downstream PIM routers on the BD may issue (*,G) Joins to different upstream routers on the BD because they have inconsistent MRIB entries regarding how to reach the RP. Both paths on the RP tree will be set up, causing two copies of all the shared tree traffic to appear on the EVPN Broadcast Domain.

* Two or more routers on the BD may issue (S,G) Joins to different upstream routers on the BD because they have inconsistent MRIB entries regarding how to reach source S. Both paths on the source-specific tree will be set up, causing two copies of all the traffic from S to appear on the BD.

* A router on the BD may issue a (*,G) Join to one upstream router on the BD, and another router on the BD may issue an (S,G) Join to a different upstream router on the same BD. Traffic from S may reach the BD over both the RPT and the SPT. If the receiver behind the downstream (*,G) router doesn't issue an (S,G,rpt) prune, then this condition would persist.

PIM does not prevent such duplicate joins from occurring; instead, when duplicate data packets appear on the same BD from different routers, these routers notice this and then elect a single forwarder. This election is performed using the PIM Assert procedure. The issue is minimized or suppressed in this document by making sure all the Upstream PEs select the same Upstream Neighbor for a given (*,G) or (S,G) in any of the three above situations. If there is only one upstream PIM router selected and the same multicast content is not allowed to be flooded from more than one Upstream Neighbor, there will not be multicast duplication or need for Assert procedures in the EVPN Broadcast Domain.

Figure 3 illustrates an example of the PIM Assert Optimization in EVPN.

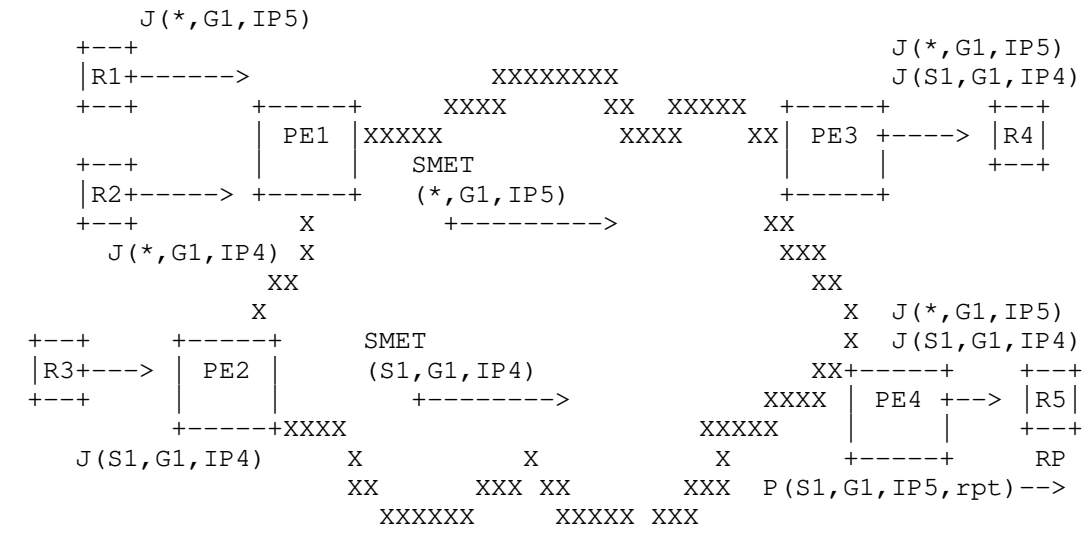

Figure 3: Proxy PIM Assert Optimization in EVPN

4.3.1. Assert Optimization Procedures in Downstream PEs

The Downstream PEs will trigger SMET routes based on the received PIM Join messages. This is their behavior when any of the three situations described in Section 4.3 occurs:

* If the Downstream PE receives two local (*,G) Joins to different Upstream Neighbors, the PE will generate a single SMET route, selecting the highest IP address. In Figure 3, if we assume R1 issues J(*,G1,IP5) and R2 J(*,G1,IP4), PE1 will advertise an SMET route for (*,G,IP5). If PE1 had already advertised (*,G1,IP4), it would have sent an update with (*,G1,IP5). Note that the Upstream Router IP address is not part of the SMET route key, hence there is no need to withdraw the previous (*,G1,IP4).

* In the same way, if the Downstream PE receives two local (S,G) Joins to different Upstream Neighbors, the PE will generate a single SMET route, selecting the highest IP address.

* If the Downstream PE receives a local (S,G) and a local (*,G) Joins for the same group but to different Upstream Neighbors, the PE will generate two different SMET routes (since *,G and S,G make two different route keys), keeping the original Upstream Neighbors in the SMET routes.

4.3.2. Assert Optimization Procedures in Upstream PEs

Upon receiving two or more SMET routes for the same group but different Upstream Neighbors, the Upstream PEs will follow this procedure:

1. The Upstream PE will select a unique Upstream Neighbor based on the following rules:

    a. The Upstream Neighbor encoded in a (S,G) SMET route has precedence over the Upstream Neighbor on the (*,G) SMET route for the same group. This is consistent with the Assert winner election in [RFC7761]. In the example of Figure 3, PE3 and PE4 will select IP4 as the Upstream Neighbor for (S1,G1) and (*,G1).

    b. In case the SMET routes have the same source (* or S), the higher Upstream Neighbor IP Address wins.

2. After selecting the Unique Upstream Neighbor, the PE will instruct the data path to discard any ingress multicast stream that is coming from an interface different than the selected Upstream Neighbor for the multicast group. In the example in Figure 3, PE4 will not accept G1 multicast traffic from R5. NOTE: when the procedure selects an Upstream Neighbor between the (S,G) and (*,G) routes, we assume that the PE's interface that is connected to the non-selected Upstream Neighbor, is not shared with another Source for the same Group. In the example of Figure 3, this means that PE4's AC cannot be shared by R5 and S2 for the same group G. If PE4's AC is connected to a switch where R5 (RP) and S2 are connected, multicast traffic (S2,G) will be dropped by PE4, as per (2).

3. Then the PE will generate the corresponding local PIM messages as usual. In the example, PE3 and PE4 generate PIM Join messages for (S1,G1,IP4) and (*,G1,IP5).

4. The PE connected to the non-selected Upstream Neighbor will issue a PIM (S,G)/(*,G) Prune or a PIM (S,G,rpt) Prune to make sure the non-selected Upstream Router does not forward traffic for the group anymore. In the example, PE4 will issue a local (S1,G1,rpt) Prune message to R5, so that R5 does not forward G1 traffic.

In case of any change that impacts on the Upstream Neighbor selection for a given group G1, the upstream PEs will simply update the Upstream Neighbor selection and follow the above procedure. This mechanism prevents the multicast duplication in the EVPN Broadcast Domain and avoids PIM Assert procedures among PIM routers in the BD.

4.4. EVPN Multi-Homing and State Synchronization

PIM Join/Prune States will be synchronized across all the PEs in an Ethernet Segment by using the procedures described in [RFC9251] and the IGMP/PIM Join Synch Route with the corresponding Flag P set. This document does not require the use of IGMP Leave Synch Routes.

In the same way, RPT-Prune States can be synchronized by using the PIM RPT-Prune Synch route. The generation and process for this route follows similar procedures as for the IGMP/PIM Join Synch Route.

In order to synchronize the PIM Neighbors discovered on an Ethernet Segment, the MRD route and its ESI value will be used. Upon receiving a Hello message on a link that is part of a multi-homed Ethernet Segment, the PE will issue an MRD route that encodes the ESI value of the AC over which the Hello was received. Upon receiving the non-zero ESI MRD route, the PEs in the same ES will add the router to their PIM Neighbor DB, using their AC on the same ES as the PIM Neighbor port. This will allow the DF on the ES to generate Hello messages for the local PIM router.

A PE that is not part of the ESI would normally receive a single non-zero ESI MRD route per multicast router. In certain transient situations the PE may receive more than one non-zero ESI MRD route for the same multicast router. The PE should recognize this and not generate additional PIM Hello messages for the local ACs.

5. Interaction with IGMP-snooping and Sources

Figure 4 illustrates an example with a multicast source, an IGMP host and a PIM router in the same EVPN BD.

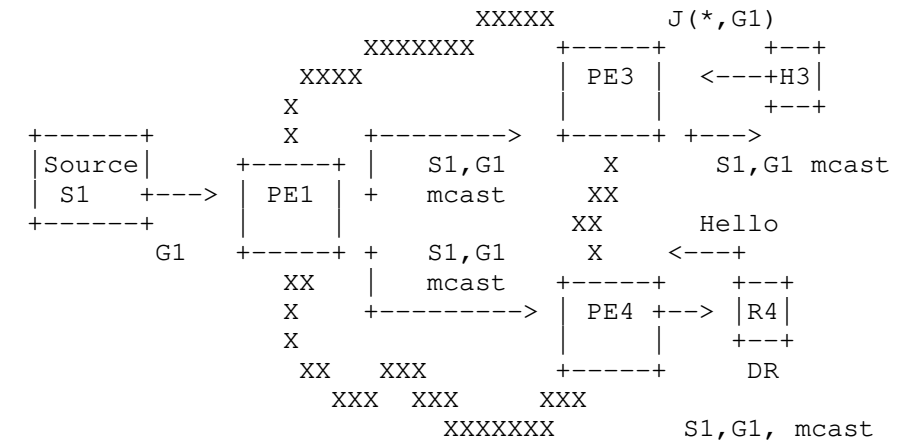

Figure 4: Proxy PIM interaction with local sources and hosts

When PIM routers, multicast sources and IGMP hosts coexist in the same EVPN Broadcast domain, the PEs supporting both IGMP and PIM proxy will provide the following optimizations in the EVPN BD:

* If an IGMP host and a PIM router are connected to the same BD on a PE, the PE will advertise a single SMET route per (S,G) or (*,G) irrespective of the received IGMP or PIM message. The IGMP flags can be simultaneously set along with the P flag.

* In the same way, if IGMP hosts and PIM routers are connected to the same BD and Ethernet Segment, the IGMP/PIM Join Synch route can be shared by a host and a router requesting the same multicast source and group.

* A PE connected to a Source and using Ingress Replication will forward a multicast stream (S1,G1) to all the egress PEs that advertised an SMET route for (S1,G1) and all the egress PEs that advertised an MRD route for the EVPN BD.

6. BGP Information Model

This document defines the following additional routes and requests IANA to allocate a type value in the EVPN route type registry:

* Type TBD - Multicast Router Discovery (MRD) Route

* Type TBD - PIM RPT-Prune Route

* Type TBD - PIM RPT-Prune Join Synch Route

In addition, the following routes defined in [RFC9251] are re-used and extended in this document's procedures:

* Type 6 - Selective Multicast Ethernet Tag Route

* Type 7 - IGMP Join Synch Route

Where Type 7 is requested to be re-named as IGMP/PIM Join Synch Route.

6.1. Multicast Router Discovery (MRD) Route

Figure 5 shows the content of the MRD route:

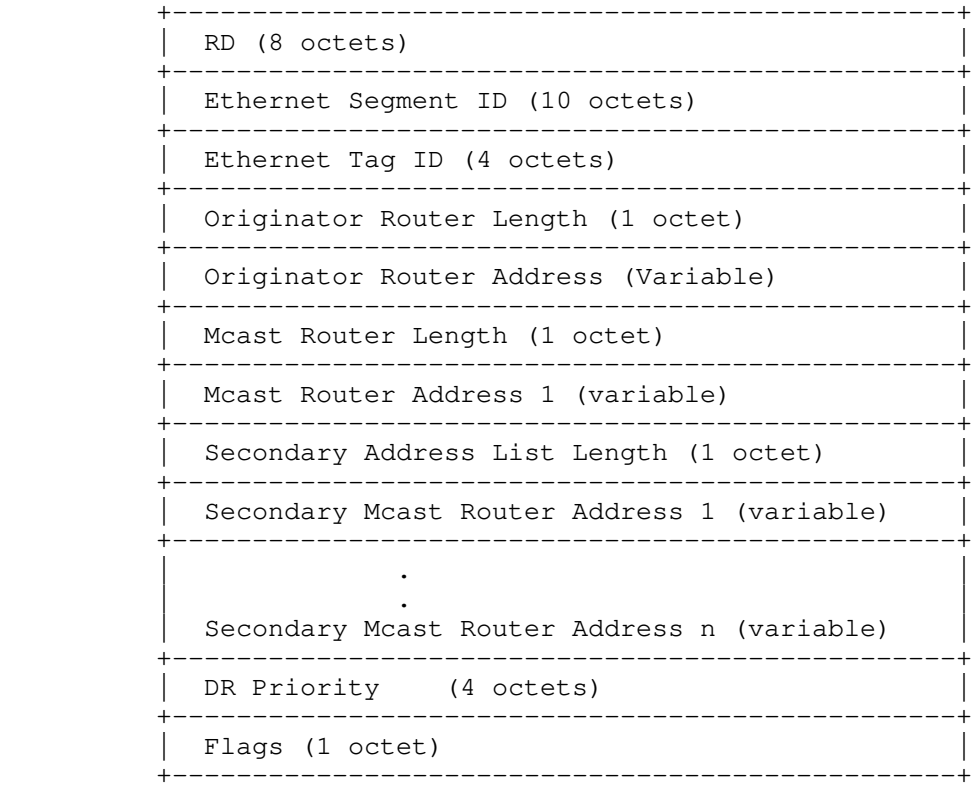

Figure 5: Multicast Router Discovery Route

The support for this new route type is OPTIONAL. Since this new route type is OPTIONAL, an implementation not supporting it MUST ignore the route, based on the unknown route type value, as specified by Section 5.4 in [RFC7606].

The encoding of this route is defined as follows:

* RD, ESI and Ethernet Tag ID are defined as per [RFC7432] for MAC/IP routes.

* The Originator Router Length and Address encode and IPv4 or IPv6 address that belongs to the advertising PE.

* The Multicast Router Length and Address field encode the Primary IP address of the PIM neighbor added to the PE's DB.

* The Secondary Address List Length encodes the number of Secondary IP addresses advertised by the PIM router in the PIM Hello message. If this field is zero, the NLRI will not include any Secondary Multicast Router Address. All the IP addresses will have the same Length, that is, they will all be either IPv4 or IPv6, but not a mix of both.

* DR Priority is copied from the same field in Hello packets, as per [RFC7761].

* Flags:

  - Q: Querier flag. Least significant bit. It indicates the encoded multicast router is an IGMP Querier.

  - P: PIM router flag. Second low order bit in the Flags octet. It indicates that the multicast router is a PIM router.

  - Q and P may be set simultaneously.

For BGP processing purposes, only the RD, Ethernet Tag ID, Originator Router Length and Address, and Multicast Router Length and Address are considered part of the route key. The Secondary Multicast Router Addresses and the rest of the fields are not part of the route key.

6.2. Selective Multicast Ethernet Tag Route for PIM Proxy

This document extends the SMET route defined in [RFC9251] as shown in Figure 6.

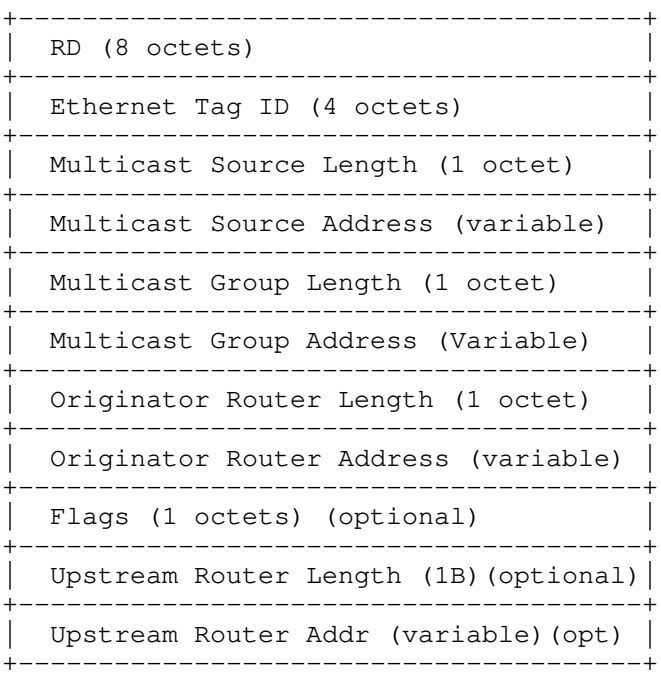

Figure 6: Selective Multicast Ethernet Tag Route and Flags

As in the case of the MRD route, this route type is OPTIONAL. This route will be used as per [RFC9251], with the following extra and optional fields:

* Upstream Router Length and Address will contain the same information as received in a PIM Join/Prune message on a local AC. There is only one Upstream Router Address per route.

* Flags: This field encodes Flags that are now relevant to IGMP and PIM. The following new Flag is defined:

- Flag P: Indicates the SMET route is generated by a received PIM Join on a local AC. When P=1, the Upstream Router Length and Address fields are present in the route. Otherwise the two fields will not be present.

Compared to [RFC9251] there is no change in terms of fields considered part of the route key for BGP processing. The Upstream Router Length and Address are not considered part of the route key.

6.3. PIM RPT-Prune Route

The RPT-Prune route is analogous to the SMET route but for PIM RPT-Prune messages. The SMET routes cannot be used to convey RPT-Prune messages because they are always triggered by IGMP or PIM Join messages. A PIM RPT-Prune message is used to Prune a specific (S,G) from the RP Tree by downstream routers. An RPT-Prune message is typically seen prior to an RPT-Join message for the (S,G), hence it requires its own BGP route type (since the SMET route is always advertised based on the received Join messages).

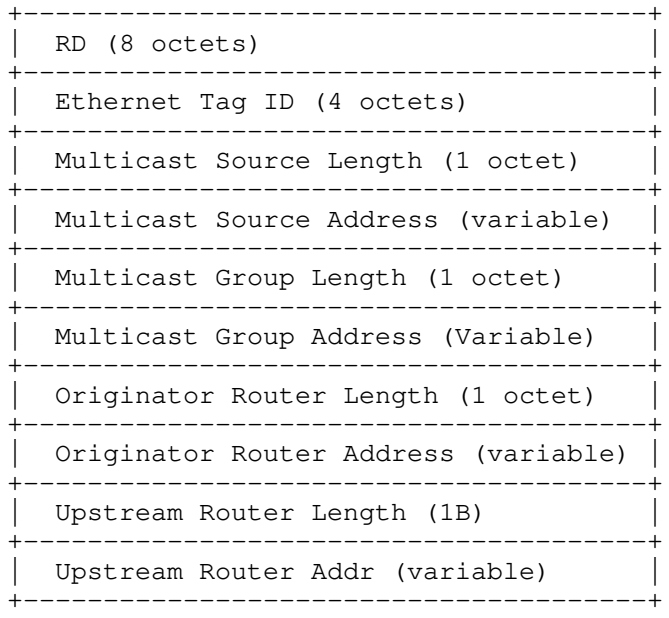

Figure 7: PIM RPT-Prune Route

Fields are defined in the same way as for the SMET route.

6.4. IGMP/PIM Join Synch Route for PIM Proxy

This document renames the IGMP Join Synch Route defined in [RFC9251] as IGMP/PIM Join Synch Route and extends it with new fields and Flags as shown in Figure 8:

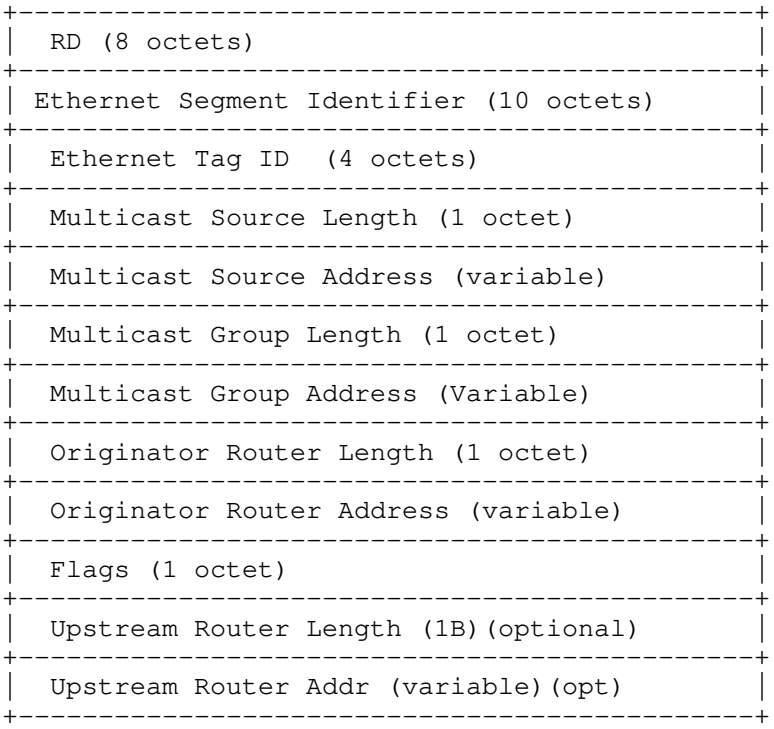

Figure 8: IGMP/PIM Join Synch Route and Flags

This route will be used as per [RFC9251], with the following extra and optional fields:

* Upstream Router Length and Address will contain the same information as received in a PIM Join/Prune message on a local AC. There is only one Upstream Router Address per route.

* Flags: This field encodes Flags that are now relevant to IGMP and PIM. The following new Flag is defined:

  - Flag P: Indicates the Join Synch route is generated by a received PIM Join on a local AC. When P=1, the Upstream Router Length and Address fields are present in the route. Otherwise the two fields will not be present.

Compared to [RFC9251] there is no change in terms of fields considered part of the route key for BGP processing. The Upstream Router Length and Address are not considered part of the route key.

6.5. IGMP/PIM RPT-Prune Synch Route for PIM Proxy

This new route is used to Synch RPT-Prune states among the PEs in the Ethernet Segment.

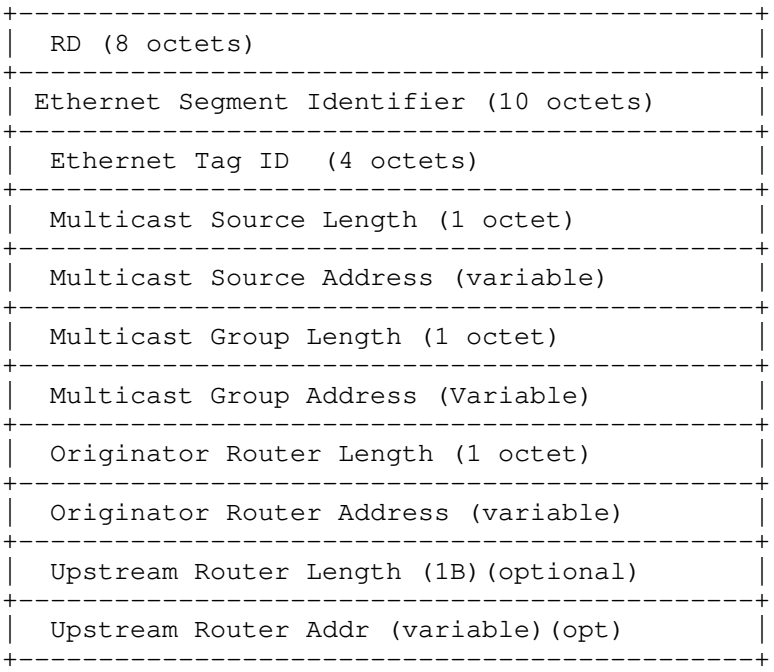

Figure 9: IGMP/PIM RPT-Prune Synch Route

The RD, Ethernet Segment Identifier and other fields are defined as for the IGMP/PIM Join Synch Route. In addition, the Upstream Router Length and Address will contain the same information as received in a PIM RPT-Prune message on a local AC. The Upstream Router points at the RP for the source and group and there is only one Upstream Router Address per route.

The route key for BGP processing is defined as per the IGMP/PIM Join Synch route.

7. Conclusions

This document extends the IGMP Proxy concept of [RFC9251] to PIM, so that EVPN can also be used to minimize the flooding of PIM control messages and optimize the delivery of IP multicast traffic in EVPN Broadcast Domains that connect PIM routers.

This specification describes procedures to Discover new PIM routers in the BD, as well as propagate PIM Join/Prune messages using EVPN SMET routes and other optimizations.

8. Security Considerations

Most of the considerations included in [RFC9251] apply to this document.

9. IANA Considerations

This document requests IANA to allocate a new EVPN route type in the corresponding registry:

* Type TBD - Multicast Router Discovery (MRD) Route

* Type TBD - PIM RPT-Prune Route

* Type TBD - PIM RPT-Prune Join Synch Route

In addition, the following route defined in [RFC9251] should be renamed as follows:

* Type 7 - IGMP/PIM Join Synch Route

10. Acknowledgments

11. Contributors

12. References

12.1. 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>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

[RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2236, DOI 10.17487/RFC2236, November 1997, <https://www.rfc-editor.org/info/rfc2236>.

[RFC8220] Dornon, O., Kotalwar, J., Hemige, V., Qiu, R., and Z. Zhang, "Protocol Independent Multicast (PIM) over Virtual Private LAN Service (VPLS)", RFC 8220, DOI 10.17487/RFC8220, September 2017, <https://www.rfc-editor.org/info/rfc8220>.

[RFC9251] Sajassi, A., Thoria, S., Mishra, M., Patel, K., Drake, J., and W. Lin, "Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Proxies for Ethernet VPN (EVPN)", RFC 9251, DOI 10.17487/RFC9251, June 2022, <https://www.rfc-editor.org/info/rfc9251>.

[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>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

12.2. Informative References

[RFC9161] Rabadan, J., Ed., Sathappan, S., Nagaraj, K., Hankins, G., and T. King, "Operational Aspects of Proxy ARP/ND in Ethernet Virtual Private Networks", RFC 9161, DOI 10.17487/RFC9161, January 2022, <https://www.rfc-editor.org/info/rfc9161>.

Authors' Addresses

Jorge Rabadan (editor)
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: jorge.rabadan@nokia.com

Jayant Kotalwar
Nokia
Email: jayant.kotalwar@nokia.com

Senthil Sathappan
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: senthil.sathappan@nokia.com

Zhaohui Zhang
Juniper Networks
United States of America
Email: zzhang@juniper.net

Ali Sajassi
Cisco Systems
822 alder drive
Milpitas, CA 95035
United States of America
Email: sajassi@cisco.com

Mankamana Mishra
Cisco Systems
822 alder drive
Milpitas, CA 95035
United States of America
Email: mankamis@cisco.com