Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 18, 2019

T. Morin, Ed.
Orange
R. Kebler, Ed.
Juniper Networks
G. Mirsky, Ed.
ZTE Corp.
February 14, 2019

Multicast VPN fast upstream failover
draft-ietf-bess-mvpn-fast-failover-05

Abstract

This document defines multicast VPN extensions and procedures that allow fast failover for upstream failures, by allowing downstream PEs to take into account the status of Provider-Tunnels (P-tunnels) when selecting the upstream PE for a VPN multicast flow, and extending BGP MVPN routing so that a C-multicast route can be advertised toward a standby upstream PE.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 18, 2019.

Table of Contents

1. Introduction 3
2. Terminology 3
3. UMH Selection based on tunnel status 3
3.1. Determining the status of a tunnel 4
3.1.1. mVPN tunnel root tracking 5
3.1.2. PE-P Upstream link status 5
3.1.3. P2MP RSVP-TE tunnels 5
3.1.4. Leaf-initiated P-tunnels 6
3.1.5. (C-S, C-G) counter information 6
3.1.6. BFD Discriminator 6
3.1.7. Per PE-CE link BFD Discriminator 9
4. Standby C-multicast route 10
4.1. Downstream PE behavior 11
4.2. Upstream PE behavior 12
4.3. Reachability determination 13
4.4. Inter-AS 13
4.4.1. Inter-AS procedures for downstream PEs, ASBR fast failover 14
4.4.2. Inter-AS procedures for ASBRs 14
5. Hot leaf standby 15
6. Duplicate packets 15
7. IANA Considerations 16
8. Security Considerations 16
9. Acknowledgments 16
10. Contributor Addresses 16
11. References 18
11.1. Normative References 18
11.2. Informative References 19
Authors' Addresses 19

1. Introduction

In the context of multicast in BGP/MPLS VPNs, it is desirable to provide mechanisms allowing fast recovery of connectivity on different types of failures. This document addresses failures of elements in the provider network that are upstream of PEs connected to VPN sites with receivers.

Section 3 describes local procedures allowing an egress PE (a PE connected to a receiver site) to take into account the status of P-tunnels to determine the Upstream Multicast Hop (UMH) for a given (C-S, C-G). This method does not provide a "fast failover" solution when used alone, but can be used with the following sections for a "fast failover" solution.

Section 4 describes protocol extensions that can speed up failover by not requiring any multicast VPN routing message exchange at recovery time.

Moreover, section 5 describes a "hot leaf standby" mechanism, that uses a combination of these two mechanisms. This approach has similarities with the solution described in [RFC7431] to improve failover times when PIM routing is used in a network given some topology and metric constraints.

2. Terminology

The terminology used in this document is the terminology defined in [RFC6513] and [RFC6514].

x-PMSI: I-PMSI or S-PMSI

3. UMH Selection based on tunnel status

Current multicast VPN specifications [RFC6513], section 5.1, describe the procedures used by a multicast VPN downstream PE to determine what the upstream multicast hop (UMH) is for a given (C-S,C-G).

The procedure described here is an OPTIONAL procedure that consists of having a downstream PE take into account the status of P-tunnels rooted at each possible upstream PEs, for including or not including each given PE in the list of candidate UMHs for a given (C-S,C-G) state. The result is that, if a P-tunnel is "down" (see Section 3.1), the PE that is the root of the P-tunnel will not be considered for UMH selection, which will result in the downstream PE to failover to the upstream PE which is next in the list of candidates. If rules to determine the state of the P-tunnel are not consistent across all PEs, then some may arrive at a different

conclusion regarding the state of the tunnel, In such a scenario, procedures described in Section 9.1.1 of [RFC6513] MUST be used.

A downstream PE monitors the status of the tunnels of UMHs that are ahead of the current one. Whenever the downstream PE determines that one of these tunnels is no longer "known to down", the PE selects the UMH corresponding to that as the new UMH.

More precisely, UMH determination for a given (C-S,C-G) will consider the UMH candidates in the following order:

- first, the UMH candidates that either (a) advertise a PMSI bound to a tunnel, where the specified tunnel is not known to be down or (b) do not advertise any x-PMSI applicable to the given (C-S,C-G) but have associated a VRF Route Import BGP attribute to the unicast VPN route for S (this is necessary to avoid incorrectly invalidating an UMH PE that would use a policy where no I-PMSI is advertised for a given VRF and where only S-PMSI are used, the S-PMSI advertisement being possibly done only after the upstream PE receives a C-multicast route for (C-S, C-G)/(C-*, C-G) to be carried over the advertised S-PMSI)

- second, the UMH candidates that advertise a PMSI bound to a tunnel that is "down" -- these will thus be used as a last resort to ensure a graceful fallback to the basic MVPN UMH selection procedures in the hypothetical case where a false negative would occur when determining the status of all tunnels

For a given downstream PE and a given VRF, the P-tunnel corresponding to a given upstream PE for a given (C-S,C-G) state is the S-PMSI tunnel advertised by that upstream PE for this (C-S,C-G) and imported into that VRF, or if there isn't any such S-PMSI, the I-PMSI tunnel advertised by that PE and imported into that VRF.

Note that this document assumes that if a site of a given MVPN that contains C-S is dual-homed to two PEs, then all the other sites of that MVPN would have two unicast VPN routes (VPN-IPv4 or VPN-IPv6) routes to C-S, each with its own RD.

3.1. Determining the status of a tunnel

Different factors can be considered to determine the "status" of a P-tunnel and are described in the following sub-sections. The optional procedures proposed in this section also allow that all downstream PEs don't apply the same rules to define what the status of a P-tunnel is (please see Section 6), and some of them will produce a result that may be different for different downstream PEs. Thus what is called the "status" of a P-tunnel in this section, is

not a characteristic of the tunnel in itself, but is the status of the tunnel, *as seen from a particular downstream PE*. Additionally, some of the following methods determine the ability of downstream PE to receive traffic on the P-tunnel and not specifically on the status of the P-tunnel itself. This could be referred to as "P-tunnel reception status", but for simplicity, we will use the terminology of P-tunnel "status" for all of these methods.

Depending on the criteria used to determine the status of a P-tunnel, there may be an interaction with another resiliency mechanism used for the P-tunnel itself, and the UMH update may happen immediately or may need to be delayed. Each particular case is covered in each separate sub-section below.

3.1.1. mVPN tunnel root tracking

A condition to consider that the status of a P-tunnel is up is that the root of the tunnel, as determined in the PMSI tunnel attribute, is reachable through unicast routing tables. In this case, the downstream PE can immediately update its UMH when the reachability condition changes.

This is similar to BGP next-hop tracking for VPN routes, except that the address considered is not the BGP next-hop address, but the root address in the PMSI tunnel attribute.

If BGP next-hop tracking is done for VPN routes and the root address of a given tunnel happens to be the same as the next-hop address in the BGP auto-discovery route advertising the tunnel, then this mechanisms may be omitted for this tunnel, as it will not bring any specific benefit.

3.1.2. PE-P Upstream link status

A condition to consider a tunnel status as Up can be that the last-hop link of the P-tunnel is up.

This method should not be used when there is a fast restoration mechanism (such as MPLS FRR [RFC4090]) in place for the link.

3.1.3. P2MP RSVP-TE tunnels

For P-tunnels of type P2MP MPLS-TE, the status of the P-tunnel is considered up if one or more of the P2MP RSVP-TE LSPs, identified by the P-tunnel Attribute, are in Up state. The determination of whether a P2MP RSVP-TE LSP is in Up state requires Path and Resv state for the LSP and is based on procedures in [RFC4875]. In this

case, the downstream PE can immediately update its UMH when the reachability condition changes.

When signaling state for a P2MP TE LSP is removed (e.g. if the ingress of the P2MP TE LSP sends a PathTear message) or the P2MP TE LSP changes state from Up to Down as determined by procedures in [RFC4875], the status of the corresponding P-tunnel SHOULD be re-evaluated. If the P-tunnel transitions from up to Down state, the upstream PE, that is the ingress of the P-tunnel, SHOULD NOT be considered a valid UMH.

3.1.4. Leaf-initiated P-tunnels

A PE can be removed from the UMH candidate list for a given (C-S, C-G) if the P-tunnel (I or S , depending) for this (S, G) is leaf triggered (PIM, mLDP), but for some reason internal to the protocol the upstream one-hop branch of the tunnel from P to PE cannot be built. In this case, the downstream PE can immediately update its UMH when the reachability condition changes.

3.1.5. (C-S, C-G) counter information

In cases, where the downstream node can be configured so that the maximum inter-packet time is known for all the multicast flows mapped on a P-tunnel, the local per-(C-S,C-G) traffic counter information for traffic received on this P-tunnel can be used to determine the status of the P-tunnel.

When such a procedure is used, in the context where fast restoration mechanisms are used for the P-tunnels, downstream PEs should be configured to wait before updating the UMH, to let the P-tunnel restoration mechanism happen. A configurable timer MUST be provided for this purpose, and it is recommended to provide a reasonable default value for this timer.

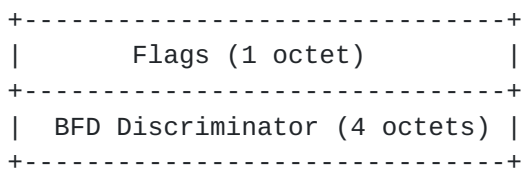
This method can be applicable, for instance, when a (C-S, C-G) flow is mapped on an S-PMSI.

In cases where this mechanism is used in conjunction with Hot leaf standby, then no prior knowledge of the rate of the multicast streams is required; downstream PEs can compare reception on the two P-tunnels to determine when one of them is down.

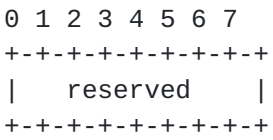
3.1.6. BFD Discriminator

P-tunnel status can be derived from the status of a multipoint BFD session [I-D.ietf-bfd-multipoint] whose discriminator is advertised along with an x-PMSI A-D route.

This document defines the format and ways of using a new BGP attribute called the "BGP- BFD attribute". This is an optional transitive BGP attribute. The format of this attribute is defined as follows:



The Flags field has the following format:



3.1.6.1. Upstream PE Procedures

When it is desired to track the P-tunnel status using p2mp BFD session, the Upstream PE:

- MUST initiate BFD session and set bfd.SessionType = MultipointHead as described in [I-D.ietf-bfd-multipoint];
- MUST use address in 127.0.0.0/8 range for IPv4 or in 0:0:0:0:0:FFFF:7F00:0/104 range for IPv6 as destination IP address when transmitting BFD control packets;
- MUST use the IP address of the Upstream PE as source IP address when transmitting BFD control packets;
- MUST include the BGP-BFD Attribute in the x-PMSI A-D Route with BFD Discriminator value set to My Discriminator value;
- MUST periodically transmit BFD control packets over the x-PMSI tunnel.

If tracking of the P-tunnel by using a p2mp BFD session is to be enabled after the P-tunnel has been already signaled, then the

procedure described above MUST be followed. Note that x-PMSI A-D Route MUST be re-sent with exactly the same attributes as before and the BGP-BFD Attribute included.

If P-tunnel is already signaled, and P-tunnel status tracked using the p2mp BFD session and it is desired to stop tracking P-tunnel status using BFD, then:

- x-PMSI A-D Route MUST be re-sent with exactly the same attributes as before, but the BGP-BFD Attribute MUST be excluded;

- the p2mp BFD session SHOULD be deleted.

3.1.6.2. Downstream PE Procedures

Upon receiving the BGP-BFD Attribute in the x-PMSI A-D Route, the Downstream PE:

- MUST associate the received BFD discriminator value with the P-tunnel originating from the Root PE and the IP address of the Upstream PE;

- MUST create p2mp BFD session and set bfd.SessionType = MultipointTail as described in [I-D.ietf-bfd-multipoint];

- MUST use the source IP address of the BFD control packet, the value of the BFD Discriminator field, and the x-PMSI tunnel identifier the BFD control packet was received to properly demultiplex BFD sessions.

After the state of the p2mp BFD session is up, i.e., bfd.SessionState == Up, the session state will then be used to track the health of the P-tunnel.

According to [I-D.ietf-bfd-multipoint], if the Downstream PE receives Down or AdminDown in the State field of the BFD control packet or associated with the BFD session Detection Timer expires, the BFD session state is down, i.e., bfd.SessionState == Down. When the BFD session state is Down, then the P-tunnel associated with the BFD session as down MUST be declared down. Then The Downstream PE MAY initiate a switchover of the traffic from the Primary Upstream PE to the Standby Upstream PE only if the Standby Upstream PE deemed available. A different p2mp BFD session MAY monitor the state of the Standby Upstream PE.

If the Downstream PE's P-tunnel is already up when the Downstream PE receives the new x-PMSI A-D Route with BGP-BFD Attribute, the Downstream PE MUST accept the x-PMSI A-D Route and associate the

value of BFD Discriminator field with the P-tunnel. The Upstream PE MUST follow procedures listed above in this section to bring the p2mp BFD session up and use it to monitor the state of the associated P-tunnel.

If the Downstream PE's P-tunnel is already up, its state being monitored by the p2mp BFD session, and the Downstream PE receives the new x-PMSI A-D Route without the BGP-BFD Attribute, the Downstream PE:

- MUST accept the x-PMSI A-D Route;
- MUST stop receiving BFD control packets for this p2mp BFD session;
- SHOULD delete the p2mp BFD session associated with the P-tunnel;
- SHOULD NOT switch the traffic to the Standby Upstream PE.

In such a scenario, in the context where fast restoration mechanisms are used for the P-tunnels, leaf PEs should be configured to wait before updating the UMH, to let the P-tunnel restoration mechanism happen. A configurable timer MUST be provided for this purpose, and it is RECOMMENDED to provide a reasonable default value for this timer.

3.1.7. Per PE-CE link BFD Discriminator

The following approach is defined for the fast failover in response to the detection of PE-CE link failures, in which UMH selection for a given C-multicast route takes into account the state of the BFD session associated with the state of the upstream PE-CE link.

3.1.7.1. Upstream PE Procedures

For each protected PE-CE link, the upstream PE initiates a multipoint BFD session [I-D.ietf-bfd-multipoint] as MultipointHead toward downstream PEs. A downstream PE monitors the state of the p2mp session as MultipointTail and MAY interpret transition of the BFD session into Down state as the indication of the associated PE-CE link being down.

For SSM groups, the upstream PE advertises an (C-S, C-G) S-PMSI A-D route or wildcard (S,*) S-PMSI A-D route for each received SSM (C-S, C-G) C-multicast route for which protection is desired. For each ASM (C-S, C-G) C-multicast route for which protection is desired, the upstream PE advertises a (C-S, C-G) S-PMSI A-D route. For each ASM (*,G) C-Multicast route for which protection is desired, the upstream PE advertises a wildcard (*,G) S-PMSI A-D route. Note that all

S-PMSI A-D routes can signal the same P-tunnel, so there is no need for a new P-tunnel for each S-PMSI A-D route. Multicast flows for which protection is desired is controlled by configuration/policy on the upstream PE. The protected link is the RPF PE-CE interface towards the src/RP. The upstream PE advertises the BFD Discriminator of the protected link in the S-PMSI A-D route. If the route to the src/RP changes such that the RPF interface is changed to be a new PE-CE interface, then the upstream PE will update the S-PMSI A-D route with included BGP-BFD Attribute so that the previously advertised value of the BFD Discriminator is associated with the new RPF link.

3.1.7.2. Downstream PE Procedures

If an S-PMSI A-D route bound to a given C-multicast is signaled with a multipoint BFD session, then the upstream PE is considered during UMH selection for the C-multicast if and only if the corresponding BFD session is not in state Down, i.e., bfd.SessionState != Down. Whenever the state of the BFD session changes to Down the Provider Tunnel will be considered down, and the downstream PE MAY switch to the backup Provider Tunnel only if the backup Provider Tunnel deemed available. The dedicated p2mp BFD session MAY monitor the state of the backup Provider Tunnel. Note that the Provider Tunnel is considered down only for the C-multicast states that match to an S-PMSI A-D route which included BGP-BFD Attribute with the BFD Discriminator of the p2mp BFD session which is down.

4. Standby C-multicast route

The procedures described below are limited to the case where the site that contains C-S is connected to two or more PEs though, to simplify the description, the case of dual-homing is described. The procedures require all the PEs of that MVPN to follow the UMH selection, as specified in [RFC6513], whether the PE selected based on its IP address, hashing algorithm described in section 5.1.3 [RFC6513], or Installed UMH Route. The procedures assume that if a site of a given MVPN that contains C-S is dual-homed to two PEs, then all the other sites of that MVPN would have two unicast VPN routes (VPN-IPv4 or VPN-IPv6) routes to C-S, each with its own RD.

As long as C-S is reachable via both PEs, a given downstream PE will select one of the PEs connected to C-S as its Upstream PE with respect to C-S. We will refer to the other PE connected to C-S as the "Standby Upstream PE". Note that if the connectivity to C-S through the Primary Upstream PE becomes unavailable, then the PE will select the Standby Upstream PE as its Upstream PE with respect to C-S. When the Primary PE later becomes available, then the PE will select the Primary Upstream PE again as its Upstream PE. This is referred to as "revertive" behavior and MUST be supported. Non-

revertive behavior would refer to the behavior of continuing to select the backup PE as the UMH even after the Primary has come up. This non-revertive behavior can also be optionally supported by an implementation and would be enabled through some configuration.

For readability, in the following sub-sections, the procedures are described for BGP C-multicast Source Tree Join routes, but they apply equally to BGP C-multicast Shared Tree Join routes failover for the case where the customer RP is dual-homed (substitute "C-RP" to "C-S").

4.1. Downstream PE behavior

When a (downstream) PE connected to some site of an MVPN needs to send a C-multicast route (C-S, C-G), then following the procedures specified in Section "Originating C-multicast routes by a PE" of [RFC6514] the PE sends the C-multicast route with RT that identifies the Upstream PE selected by the PE originating the route. As long as C-S is reachable via the Primary Upstream PE, the Upstream PE is the Primary Upstream PE. If C-S is reachable only via the Standby Upstream PE, then the Upstream PE is the Standby Upstream PE.

If C-S is reachable via both the Primary and the Standby Upstream PE, then in addition to sending the C-multicast route with an RT that identifies the Primary Upstream PE, the PE also originates and sends a C-multicast route with an RT that identifies the Standby Upstream PE. This route, that has the semantics of being a 'standby' C-multicast route, is further called a "Standby BGP C-multicast route", and is constructed as follows:

- the NLRI is constructed as the original C-multicast route, except that the RD is the same as if the C-multicast route was built using the standby PE as the UMH (it will carry the RD associated to the unicast VPN route advertised by the standby PE for S and a Route Target derived from the standby PE's UMH route's VRF RT Import EC);

- SHOULD carry the "Standby PE" BGP Community (this is a new BGP Community, see Section 7).

The normal and the standby C-multicast routes must have their Local Preference attribute adjusted so that, if two C-multicast routes with same NLRI are received by a BGP peer, one carrying the "Standby PE" attribute and the other one *not* carrying the "Standby PE" community, then preference is given to the one *not* carrying the "Standby PE" attribute. Such a situation can happen when, for instance, due to transient unicast routing inconsistencies, two different downstream PEs consider different upstream PEs to be the

primary one; in that case, without any precaution taken, both upstream PEs would process a standby C-multicast route and possibly stop forwarding at the same time. For this purpose, routes that carry the "Standby PE" BGP Community MUST have the LOCAL_PREF attribute set to zero.

Note that, when a PE advertises such a Standby C-multicast join for an (C-S, C-G) it must join the corresponding P-tunnel.

If at some later point the local PE determines that C-S is no longer reachable through the Primary Upstream PE, the Standby Upstream PE becomes the Upstream PE, and the local PE re-sends the C-multicast route with RT that identifies the Standby Upstream PE, except that now the route does not carry the Standby PE BGP Community (which results in replacing the old route with a new route, with the only difference between these routes being the presence/absence of the Standby PE BGP Community).

4.2. Upstream PE behavior

When a PE receives a C-multicast route for a particular (C-S, C-G), and the RT carried in the route results in importing the route into a particular VRF on the PE, if the route carries the Standby PE BGP Community, then the PE performs as follows:

> when the PE determines that C-S is not reachable through some other PE, the PE SHOULD install VRF PIM state corresponding to this Standby BGP C-multicast route (the result will be that a PIM Join message will be sent to the CE towards C-S, and that the PE will receive (C-S,C-G) traffic), and the PE SHOULD forward (C-S, C-G) traffic received by the PE to other PEs through a P-tunnel rooted at the PE.

Furthermore, irrespective of whether C-S carried in that route is reachable through some other PE:

a) based on local policy, as soon as the PE receives this Standby BGP C-multicast route, the PE MAY install VRF PIM state corresponding to this BGP Source Tree Join route (the result will be that Join messages will be sent to the CE toward C-S, and that the PE will receive (C-S,C-G) traffic)

b) based on local policy, as soon as the PE receives this Standby BGP C-multicast route, the PE MAY forward (C-S, C-G) traffic to other PEs through a P-tunnel independently of the reachability of C-S through some other PE. [note that this implies also doing (a)]

Doing neither (a) or (b) for a given (C-S,C-G) is called "cold root standby".

Doing (a) but not (b) for a given (C-S,C-G) is called "warm root standby".

Doing (b) (which implies also doing (a)) for a given (C-S,C-G) is called "hot root standby".

Note that, if an upstream PE uses an S-PMSI only policy, it shall advertise an S-PMSI for an (C-S, C-G) as soon as it receives a C-multicast route for (C-S, C-G), normal or Standby; i.e., it shall not wait for receiving a non-Standby C-multicast route before advertising the corresponding S-PMSI.

Section 9.3.2 of [RFC6514], describes the procedures of sending a Source-Active A-D result as a result of receiving the C-multicast route. These procedures should be followed for both the normal and Standby C-multicast routes.

4.3. Reachability determination

The standby PE can use the following information to determine that C-S can or cannot be reached through the primary PE:

- presence/absence of a unicast VPN route toward C-S
- supposing that the standby PE is an egress of the tunnel rooted at the Primary PE, the standby PE can determine the reachability of C-S through the Primary PE based on the status of this tunnel, determined thanks to the same criteria as the ones described in Section 3.1 (without using the UMH selection procedures of Section 3);
- other mechanisms MAY be used.

4.4. Inter-AS

If the non-segmented inter-AS approach is used, the procedures in section 4 can be applied.

When multicast VPNs are used in an inter-AS context with the segmented inter-AS approach described in section 8.2 of [RFC6514], the procedures in this section can be applied.

A pre-requisite for the procedures described below to be applied for a source of a given MVPN is:

- that any PE of this MVPN receives two Inter-AS I-PMSI auto-discovery routes advertised by the AS of the source (or more)

- that these Inter-AS I-PMSI auto-discovery routes have distinct Route Distinguishers (as described in item "(2)" of section 9.2 of [RFC6514]).

As an example, these conditions will be satisfied when the source is dual-homed to an AS that connects to the receiver AS through two ASBR using auto-configured RDs.

4.4.1. Inter-AS procedures for downstream PEs, ASBR fast failover

The following procedure is applied by downstream PEs of an AS, for a source S in a remote AS.

Additionally, to choosing an Inter-AS I-PMSI auto-discovery route advertised from the AS of the source to construct a C-multicast route, as described in section 11.1.3 [RFC6514] a downstream PE will choose a second Inter-AS I-PMSI auto-discovery route advertised from the AS of the source and use this route to construct and advertise a Standby C-multicast route (C-multicast route carrying the Standby extended community) as described in Section 4.1.

4.4.2. Inter-AS procedures for ASBRs

When an upstream ASBR receives a C-multicast route, and at least one of the RTs of the route matches one of the ASBR Import RT, the ASBR locates an Inter-AS I-PMSI A-D route whose RD and Source AS matches the RD and Source AS carried in the C-multicast route. If the match is found, and C-multicast route carries the Standby PE BGP Community, then the ASBR performs as follows:

- if the route was received over iBGP; the route is expected to have a LOCAL_PREF attribute set to zero and it should be re-advertised in eBGP with a MED attribute (MULTI_EXIT_DISC) set to the highest possible value (0xffff)

- if the route was received over eBGP; the route is expected to have a MED attribute set of 0xffff and should be re-advertised in iBGP with a LOCAL_PREF attribute set to zero

Other ASBR procedures are applied without modification.

5. Hot leaf standby

The mechanisms defined in sections Section 4 and Section 3 can be used together as follows.

The principle is that, for a given VRF (or possibly only for a given C-S,C-G):

- downstream PEs advertise a Standby BGP C-multicast route (based on Section 4)
- upstream PEs use the "hot standby" optional behavior and thus will forward traffic for a given multicast state as soon as they have whether a (primary) BGP C-multicast route or a Standby BGP C-multicast route for that state (or both)
- downstream PEs accept traffic from the primary or standby tunnel, based on the status of the tunnel (based on Section 3)

Other combinations of the mechanisms proposed in Section 4) and Section 3 are for further study.

Note that the same level of protection would be achievable with a simple C-multicast Source Tree Join route advertised to both the primary and secondary upstream PEs (carrying as Route Target extended communities, the values of the VRF Route Import attribute of each VPN route from each upstream PEs). The advantage of using the Standby semantic for is that, supposing that downstream PEs always advertise a Standby C-multicast route to the secondary upstream PE, it allows to choose the protection level through a change of configuration on the secondary upstream PE, without requiring any reconfiguration of all the downstream PEs.

6. Duplicate packets

Multicast VPN specifications [RFC6513] impose that a PE only forwards to CEs the packets coming from the expected upstream PE (Section 9.1).

We highlight the reader's attention to the fact that the respect of this part of multicast VPN specifications is especially important when two distinct upstream PEs are susceptible to forward the same traffic on P-tunnels at the same time in the steady state. This will be the case when "hot root standby" mode is used (Section 4), and which can also be the case if procedures of Section 3 are used and (a) the rules determining the status of a tree are not the same on two distinct downstream PEs or (b) the rule determining the status of

a tree depend on conditions local to a PE (e.g. the PE-P upstream link being up).

7. IANA Considerations

Allocation is expected from IANA for the BGP "Standby PE" community. (TBC)

8. Security Considerations

9. Acknowledgments

The authors want to thank Greg Reaume, Eric Rosen, Jeffrey Zhang, and Zheng (Sandy) Zhang for their reviews, useful comments, and helpful suggestions.

10. Contributor Addresses

Below is a list of other contributing authors in alphabetical order:

Rahul Aggarwal
Arktan

Email: raggarwa_1@yahoo.com

Nehal Bhau
Alcatel-Lucent, Inc.
701 E Middlefield Rd
Mountain View, CA 94043
USA

Email: Nehal.Bhau@alcatel-lucent.com

Clayton Hassen
Bell Canada
2955 Virtual Way
Vancouver
CANADA

Email: Clayton.Hassen@bell.ca

Wim Henderickx

Alcatel-Lucent
Copernicuslaan 50
Antwerp 2018
Belgium

Email: wim.henderickx@alcatel-lucent.com

Pradeep Jain
Alcatel-Lucent, Inc.
701 E Middlefield Rd
Mountain View, CA 94043
USA

Email: pradeep.jain@alcatel-lucent.com

Jayant Kotalwar
Alcatel-Lucent, Inc.
701 E Middlefield Rd
Mountain View, CA 94043
USA

Email: Jayant.Kotalwar@alcatel-lucent.com

Praveen Muley
Alcatel-Lucent
701 East Middlefield Rd
Mountain View, CA 94043
U.S.A.

Email: praveen.muley@alcatel-lucent.com

Ray (Lei) Qiu
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
U.S.A.

Email: rqiu@juniper.net

Yakov Rekhter
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
U.S.A.

Email: yakov@juniper.net

Kanwar Singh
Alcatel-Lucent, Inc.
701 E Middlefield Rd
Mountain View, CA 94043
USA

Email: kanwar.singh@alcatel-lucent.com

11. References

11.1. Normative References

[I-D.ietf-bfd-multipoint]
Katz, D., Ward, D., Networks, J., and G. Mirsky, "BFD for Multipoint Networks", draft-ietf-bfd-multipoint-19 (work in progress), December 2018.

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

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

[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>.

[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>.

11.2. Informative References

[RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, DOI 10.17487/RFC4090, May 2005, <https://www.rfc-editor.org/info/rfc4090>.

[RFC7431] Karan, A., Filsfils, C., Wijnands, IJ., Ed., and B. Decraene, "Multicast-Only Fast Reroute", RFC 7431, DOI 10.17487/RFC7431, August 2015, <https://www.rfc-editor.org/info/rfc7431>.

Authors' Addresses

Thomas Morin (editor)
Orange
2, avenue Pierre Marzin
Lannion 22307
France

Email: thomas.morin@orange-ftgroup.com

Robert Kebler (editor)
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
U.S.A.

Email: rkebler@juniper.net

Greg Mirsky (editor)
ZTE Corp.

Email: gregimirsky@gmail.com