Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 1, 2021

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

October 28, 2020

Multicast VPN Fast Upstream Failover
draft-ietf-bess-mvpn-fast-failover-12

Abstract

This document defines multicast VPN extensions and procedures that allow fast failover for upstream failures by allowing downstream PEs to consider the status of Provider-Tunnels (P-tunnels) when selecting the upstream PE for a VPN multicast flow. The fast failover is enabled by using RFC 8562 BFD for Multipoint Networks and the new BGP Attribute - BFD Discriminator. Also, the document introduces a new BGP Community, Standby PE, extending BGP MVPN routing so that a C-multicast route can be advertised toward a Standby Upstream PE.

Table of Contents

1. Introduction

It is assumed that the reader is familiar with the workings of multicast MPLS/BGP IP VPNs as described in [RFC6513] and [RFC6514].

In the context of multicast in BGP/MPLS VPNs [RFC6513], 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). One of the optional methods uses [RFC8562] and the new BGP Attribute - BFD Discriminator. None of these methods provide a "fast failover" solution when used alone, but can be used together with the mechanism described in Section 4 for a "fast failover" solution.

Section 4 describes an optional BGP extension, a new Standby PE Community. that can speed up failover by not requiring any multicast VPN routing message exchange at recovery time.

Section 5 describes a "hot leaf standby" mechanism that can be used to improve failover time in MVPN. The approach combines mechanisms defined in Section 3 and Section 4 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.

The procedures described in this document are optional to enable an operator to provide protection for multicast services in BGP/MPLS IP VPNs. An operator would enable these mechanisms using a method discussed in Section 3 in combination with the redundancy provided by a standby PE connected to the source of the multicast flow, and it is assumed that all PEs in the network would support these mechanisms for the procedures to work. In the case that a BGP implementation does not recognize or is configured to not support the extensions defined in this document, it will continue to provide the multicast service, as described in [RFC6513].

2. Conventions used in this document

2.1. 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.

2.2. Terminology

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

The term 'upstream' (lower case) throughout this document refers to links and nodes that are upstream to a PE connected to VPN sites with receivers of a multicast flow.

The term 'Upstream' (capitalized) throughout this document refers to a PE or an Autonomous System Border Router (ASBR) at which (S,G) or (*,G) data packets enter the VPN backbone or the local AS when traveling through the VPN backbone.

2.3. Acronyms

PMSI: P-Multicast Service Interface

I-PMSI: Inclusive PMSI

S-PMSI: Selective PMSI

x-PMSI: Either an I-PMSI or an S-PMSI

P-tunnel: Provider-Tunnels

UMH: Upstream Multicast Hop

VPN: Virtual Private Network

MVPN: Multicast VPN

RD: Route Distinguisher

RP: Rendezvous Point

NLRI: Network Layer Reachability Information

VRF: VPN Routing and Forwarding Table

MED: Multi-Exit Discriminator

P2MP: Point-to-Multipoint

3. UMH Selection Based on Tunnel Status

Section 5.1 of [RFC6513] describes procedures used by a multicast VPN downstream PE to determine the Upstream Multicast Hop (UMH) for a given (C-S, C-G).

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.

The procedure described here is an OPTIONAL procedure that is based on a downstream PE taking into account the status of P-tunnels rooted at each possible Upstream PE, for including or not including each given PE in the list of candidate UMHs for a given (C-S, C-G) state. If it is not possible to determine whether a P-tunnel's current status is Up, the state shall be considered "not known to be Down", and it may be treated as if it is Up so that attempts to use the tunnel are acceptable. 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. This will result in the downstream PE failing over to use the next Upstream PE in the list of candidates. Some downstream PEs could arrive at a different conclusion regarding the tunnel's state because the failure impacts only a subset of branches. Because of that, the procedures of Section 9.1.1 of [RFC6513] are applicable when using I-PMSI P-tunnels. That document is a foundation for this document, and its processes all apply here. Section 9.1.1 mandates the use of specific procedures for sending intra-AS I-PMSI A-D Routes.

There are three options specified in Section 5.1 of [RFC6513] for a downstream PE to select an Upstream PE.

- The first two options select the Upstream PE from a candidate PE set either based on an IP address or a hashing algorithm. When used together with the optional procedure of considering the P-tunnel status as in this document, a candidate Upstream PE is included in the set if it either:

  A. advertises an x-PMSI bound to a tunnel, where the specified tunnel's state is not known to be Down, or,

  B. does not advertise any x-PMSI applicable to the given (C-S, C-G) but has associated a VRF Route Import BGP attribute to the unicast VPN route for S. That is necessary to avoid incorrectly invalidating a 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 can be advertised 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.

If the resulting candidate set is empty, then the procedure is repeated without considering the P-tunnel status.

o The third option uses the installed UMH Route (i.e., the "best" route towards the C-root) as the Selected UMH Route, and its originating PE is the selected Upstream PE. With the optional procedure of considering P-tunnel status as in this document, the Selected UMH Route is the best one among those whose originating PE's P-tunnel is not "down". If that does not exist, the installed UMH Route is selected regardless of the P-tunnel status.

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 described in this section also handle the case the downstream PEs do not all 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, the "status" of a P-tunnel in this section is not a characteristic of the tunnel in itself, but is the tunnel status, as seen from a particular downstream PE. Additionally, some of the following methods determine the ability of a downstream PE to receive traffic on the P-tunnel and not specifically on the status of the P-tunnel itself. That 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.

An implementation may support any combination of the methods described in this section and provide a network operator with control to choose which one to use in the particular deployment.

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 x-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.

That 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 x-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 A-D Route advertising the tunnel, then checking, in unicast routing tables, whether the tunnel root is reachable, will be unnecessary duplication and thus 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. Conversely, if the last-hop link of the P-tunnel is Down then this can be taken as an indication that the P-tunnel is Down.

Using this method when a fast restoration mechanism (such as MPLS FRR [RFC4090]) is in place for the link requires careful consideration and coordination of defect detection intervals for the link and the tunnel. In many cases, it is not practical to use both protection methods at the same time because uncorrelated timers might cause unnecessary switchovers and destabilize the network.

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 the sub-LSP to this downstream PE is in the Up state. The determination of whether a P2MP RSVP-TE LSP is in the Up state requires Path and Resv state for the LSP and is based on procedures specified in [RFC4875]. As a result, the downstream PE can immediately update its UMH when the reachability condition changes.

When using this method and if the 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 MUST be re-evaluated. If the P-tunnel transitions from Up to Down state, the Upstream PE that is the ingress of the P-tunnel MUST NOT be considered a valid UMH.

3.1.4. Leaf-initiated P-tunnels

An Upstream PE SHOULD be removed from the UMH candidate list for a given (C-S, C-G) if the P-tunnel (I-PMSI or S-PMSI) 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. As a result, 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, a configurable timer MUST be set on the downstream PE to wait before updating the UMH, to let the P-tunnel restoration mechanism to execute its actions. An implementation SHOULD use three seconds as the default value for this timer.

In cases where this mechanism is used in conjunction with the method described in Section 5, 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 Attribute

P-tunnel status may be derived from the status of a multipoint BFD session [RFC8562] 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 "BFD Discriminator". It is an optional transitive BGP attribute. An implementation that does not recognize or is configured not to support this attribute MUST follow procedures defined for optional transitive path attributes in Section 5 of [RFC4271]. In Section 7.2, IANA is requested to allocate the codepoint value (TBA2). The format of this attribute is shown in Figure 1.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    BFD Mode   |                  Reserved                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       BFD Discriminator                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                         Optional TLVs                         ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: Format of the BFD Discriminator Attribute

Where:

BFD Mode field is the one octet long. This specification defines the P2MP BFD Session as value 1 Section 7.2.

Reserved field is three octets long, and the value MUST be zeroed on transmission and ignored on receipt.

BFD Discriminator field is four octets long.

Optional TLVs is the optional variable-length field that MAY be used in the BFD Discriminator attribute for future extensions. TLVs MAY be included in a sequential or nested manner. To allow for TLV nesting, it is advised to define a new TLV as a variable-length object. Figure 2 presents the Optional TLV format TLV that consists of:

* one octet-long field of TLV's Type value (Section 7.3)

* one octet-long field of the length of the Value field in octets

* variable length Value field.

The length of a TLV MUST be multiple of four octets.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type     |     Length    |           Value             ...
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Format of the Optional TLV

The BFD Discriminator attribute MUST be considered malformed if its length is not a non-zero multiple of four. If the attribute considered malformed, the UPDATE message SHALL be handled using the approach of Attribute Discard per [RFC7606].

3.1.6.1. Upstream PE Procedures

To enable downstream PEs to track the P-tunnel status using a point-to-multipoint (P2MP) BFD session the Upstream PE:

- MUST initiate the BFD session and set bfd.SessionType = MultipointHead as described in [RFC8562];

- MUST set the IP destination address of the inner IP header to one of the internal loopback addresses from 127/8 range for IPv4 or one of IPv4-mapped IPv6 addresses from ::ffff:127.0.0.0/104 range for IPv6 when transmitting BFD Control packets;

- MUST use its IP address as the source IP address when transmitting BFD Control packets;

- MUST include the BFD Discriminator attribute in the x-PMSI A-D Route with the value set to My Discriminator value;

- MUST periodically transmit BFD Control packets over the x-PMSI P-tunnel after the P-tunnel is considered established. Note that the methods to declare a P-tunnel has been established are outside the scope of this specification.

If the tracking of the P-tunnel by using a P2MP BFD session is enabled after the x-PMSI A-D Route has been already advertised, the x-PMSI A-D Route MUST be re-sent with precisely the same attributes as before and the BFD Discriminator attribute included.

If the x-PMSI A-D Route is advertised with 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 precisely the same attributes as before, but the BFD Discriminator attribute MUST be excluded;

- the P2MP BFD session SHOULD be deleted.

3.1.6.2. Downstream PE Procedures

Upon receiving the BFD Discriminator 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 Upstream PE and the IP address of the Upstream PE;

- MUST create a P2MP BFD session and set bfd.SessionType = MultipointTail as described in [RFC8562];

- MUST use the source IP address of the BFD Control packet, the value of the BFD Discriminator field, and the x-PMSI Tunnel Identifier [RFC6514] 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 [RFC8562], 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 is down, i.e., bfd.SessionState == Down. When the BFD session state is Down, then the P-tunnel associated with the BFD session MUST be considered down. If the site that contains C-S is connected to two or more PEs, a downstream PE will select one as its Primary Upstream PE, while others are considered as Standby Upstream PEs. In such a scenario, when the P-tunnel is considered down, 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 is deemed available.

If the downstream PE's P-tunnel is already established when the downstream PE receives the new x-PMSI A-D Route with BFD Discriminator attribute, the downstream PE MUST associate the value of BFD Discriminator field with the P-tunnel and follow procedures listed above in this section if and only if the x-PMSI A-D Route was properly processed as per [RFC6514], and the BFD Discriminator attribute was validated.

If the downstream PE's P-tunnel is already established, its state being monitored by the P2MP BFD session, and the downstream PE receives the new x-PMSI A-D Route without the BFD Discriminator attribute, and the x-PMSI A-D Route was processed without any error as per the relevant specifications, the downstream PE:

- o MUST stop processing BFD Control packets for this P2MP BFD session;
- o SHOULD delete the P2MP BFD session associated with the P-tunnel;
- o SHOULD NOT switch the traffic to the Standby Upstream PE.

3.1.7. Per PE-CE Link BFD Discriminator

The following approach is defined in response to the detection by the Upstream PE of a PE-CE link failure. Even though the provider tunnel is still up, it is desired for the downstream PEs to switch to a backup Upstream PE. To achieve that, if the Upstream PE detects that its PE-CE link fails, it SHOULD set the bfd.LocalDiag of the P2MP BFD session to Concatenated Path Down and/or Reverse Concatenated Path Down (per Section 6.8.17 [RFC5880]), unless it switches to a new PE-CE link within the time of bfd.DesiredMinTxInterval for the P2MP BFD session (in that case, the Upstream PE will start tracking the status of the new PE-CE link). When a downstream PE receives that bfd.LocalDiag code, it treats it as if the tunnel itself failed and tries to switch to a backup PE.

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 same UMH selection procedure, as specified in [RFC6513], whether the PE selected based on its IP address, hashing algorithm described in section 5.1.3 of [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) to C-S, each with its 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 for 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 for C-S. When the Primary PE later becomes available, then the PE will select the Primary Upstream PE again as its Upstream PE. Such behavior is referred to as "revertive" behavior and MUST be supported. Non-revertive behavior refers to the behavior of continuing to select the backup PE as the UMH even after the Primary has come up. This non-revertive behavior MAY also be 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 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 11.1 of [RFC6514], the PE sends the C-multicast route with an 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 downstream PE also originates and sends a C-multicast route with an RT that identifies the Standby Upstream PE. The 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:

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

- o MUST carry the "Standby PE" BGP Community (this is a new BGP Community. Section 7.1 requested IANA to allocate value TBA1).

The Local Preference attribute of the normal and the standby C-multicast route needs to be adjusted. so that, if a BGP peer receives two C-multicast routes with the same NLRI, one carrying the "Standby PE" community and the other one not carrying the "Standby PE" community, then preference is given to the one not carrying the "Standby PE" community. Such a situation can happen when, for instance, due to transient unicast routing inconsistencies or lack of support of the Standby PE community, 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 a (C-S, C-G) it MUST join the corresponding P-tunnel.

If at some later point, the PE determines that C-S is no longer reachable through the Primary Upstream PE, the Standby Upstream PE becomes the Upstream PE, and the 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). The LOCAL_PREF attribute MUST be set to zero.

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 (the use of the particular method to detect the failure is outside the scope of this document) 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 a (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 Route as a result of receiving the C-multicast route. These procedures MUST be followed for both the normal and Standby C-multicast routes.

4.3. Reachability Determination

The Standby Upstream PE can use the following information to determine that C-S can or cannot be reached through the Primary Upstream PE:

- o presence/absence of a unicast VPN route toward C-S

- o supposing that the Standby Upstream PE is the egress of the tunnel rooted at the Primary Upstream PE, the Standby Upstream PE can determine the reachability of C-S through the Primary Upstream 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);

- o other mechanisms MAY be used.

4.4. Inter-AS

If the non-segmented inter-AS approach is used, the procedures described in Section 4.1 through Section 4.3 can be applied.

When multicast VPNs are used in an inter-AS context with the segmented inter-AS approach described in Section 9.2 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:

- o that any PE of this MVPN receives two or more Inter-AS I-PMSI A-D Routes advertised by the AS of the source

- o that these Inter-AS I-PMSI A-D 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 A-D 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 A-D 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, that supports this specification, MUST try to locate an Inter-AS I-PMSI A-D Route whose RD and Source AS respectively match the RD and Source AS carried in the C-multicast route. If the match is found, and the C-multicast route carries the Standby PE BGP Community, then the ASBR MUST perform as follows:

- o if the route was received over iBGP and its LOCAL_PREF attribute is set to zero, then it MUST be re-advertised in eBGP with a MED attribute (MULTI_EXIT_DISC) set to the highest possible value (0xffff)

- o if the route was received over eBGP and its MED attribute set to 0xffff, then it MUST be re-advertised in iBGP with a LOCAL_PREF attribute set to zero

Other ASBR procedures are applied without modification.

5. Hot Root Standby

The mechanisms defined in 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):

- o downstream PEs advertise a Standby BGP C-multicast route (based on Section 4)

- o 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)

- o 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 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 of [RFC6513]).

We draw 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. That 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 depends on conditions local to a PE (e.g., the PE-P upstream link being up).

7. IANA Considerations

7.1. Standby PE Community

IANA is requested to allocate the BGP "Standby PE" community value (TBA1) from the Border Gateway Protocol (BGP) Well-known Communities registry using the First Come First Served registration policy.

7.2. BFD Discriminator

This document defines a new BGP optional transitive attribute, called "BFD Discriminator". IANA is requested to allocate a codepoint (TBA2) in the "BGP Path Attributes" registry to the BFD Discriminator attribute.

IANA is requested to create a new BFD Mode sub-registry in the Border Gateway Protocol (BGP) Parameters registry. The registration policies, per [RFC8126], for this sub-registry are according to Table 1.

Value	Policy
0- 175	IETF Review
176 - 249	First Come First Served
250 - 254	Experimental Use
255	IETF Review

Table 1: BFD Mode Sub-registry Registration Policies

IANA is requested to make initial assignments according to Table 2.

Value	Description	Reference
0	Reserved	This document
1	P2MP BFD Session	This document
2- 175	Unassigned	This document
176 - 249	Unassigned	This document
250 - 254	Experimental Use	This document
255	Reserved	This document

Table 2: BFD Mode Sub-registry

7.3. BFD Discriminator Optional Sub-TLV Type

IANA is requested to create a new BFD Discriminator Optional sub-TLV Type sub-registry in Border Gateway Protocol (BGP). The registration policies, per [RFC8126], for this sub-registry are according to Table 3.

Value	Policy
0- 175	IETF Review
176 - 249	First Come First Served
250 - 254	Experimental Use
255	IETF Review

Table 3: BFD Discriminator Optional Sub-TLV Type Sub-registry Registration Policies

IANA is requested to make initial assignments according to Table 4.

Value	Description	Reference
0	Reserved	This document
1- 175	Unassigned	This document
176 - 249	Unassigned	This document
250 - 254	Experimental Use	This document
255	Reserved	This document

Table 4: BFD Discriminator Optional Sub-TLV Type Sub-registry

8. Security Considerations

This document describes procedures based on [RFC6513] and [RFC6514] and hence shares the security considerations respectively represented in these specifications.

This document uses P2MP BFD, as defined in [RFC8562], which, in turn, is based on [RFC5880]. Security considerations relevant to each protocol are discussed in the respective protocol specifications. An implementation that supports this specification MUST use a mechanism to control the maximum number of P2MP BFD sessions that can be active at the same time.

9. Acknowledgments

The authors want to thank Greg Reaume, Eric Rosen, Jeffrey Zhang, Martin Vigoureux, Adrian Farrel, 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
Cisco

Email: NBhau@cisco.com

Clayton Hassen
Bell Canada
2955 Virtual Way
Vancouver
CANADA

Email: Clayton.Hassen@bell.ca

Wim Henderickx
Nokia
Copernicuslaan 50
Antwerp 2018
Belgium

Email: wim.henderickx@nokia.com

Pradeep Jain
Nokia
701 E Middlefield Rd
Mountain View, CA 94043

USA

Email: pradeep.jain@nokia.com

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

Email: Jayant.Kotalwar@nokia.com

Praveen Muley
Nokia
701 East Middlefield Rd
Mountain View, CA 94043
U.S.A.

Email: praveen.muley@nokia.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
Nokia
701 E Middlefield Rd

Mountain View, CA 94043
USA

Email: kanwar.singh@nokia.com

11. References

11.1. Normative References

[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>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[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>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <https://www.rfc-editor.org/info/rfc5880>.

[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>.

[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>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[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>.

[RFC8562] Katz, D., Ward, D., Pallagatti, S., Ed., and G. Mirsky, Ed., "Bidirectional Forwarding Detection (BFD) for Multipoint Networks", RFC 8562, DOI 10.17487/RFC8562, April 2019, <https://www.rfc-editor.org/info/rfc8562>.

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