Internet Engineering Task Force (IETF)
Request for Comments: 9026
Category: Standards Track
ISSN: 2070-1721

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

April 2021

Multicast VPN Fast Upstream Failover

Abstract

This document defines Multicast Virtual Private Network (VPN) extensions and procedures that allow fast failover for upstream failures by allowing downstream Provider Edges (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 "Bidirectional Forwarding Detection (BFD) for Multipoint Networks" (RFC 8562) and the new BGP Attribute, BFD Discriminator. Also, this document introduces a new BGP Community, Standby PE, extending BGP Multicast VPN (MVPN) routing so that a C-multicast route can be advertised toward a Standby Upstream PE.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9026.

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 (MVPN) routing message exchange at recovery time.

Section 5 describes a "hot root standby" mechanism that can be used to improve failover time in MVPN. The approach combines mechanisms defined in Sections 3 and 4 and 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 and allow 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 combined with the redundancy provided by a standby PE connected to the multicast flow source. PEs that support these mechanisms would converge faster and thus provide a more stable multicast service. In the case that a BGP implementation does not recognize or is configured not to support the extensions defined in this document, the implementation 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. Abbreviations

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-Tunnel

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 an MVPN 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 that (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 optional one, 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.

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 based either 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 Extended Community 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-PMSIs 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.

* 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 subsections. The optional procedures described in this section also handle the case when 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 subsection 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

When determining if the status of a P-tunnel is Up, a condition to

consider is whether the root of the tunnel, as specified 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. BGP next-hop tracking monitors BGP next-hop address changes in the routing table. In general, when a change is detected, it performs a next-hop scan to find if any of the next hops in the BGP table is affected and updates it accordingly.

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 will thus not bring any specific benefit.

3.1.2. PE-P Upstream Link Status

When determining if the status of a P-tunnel is Up, a condition to consider is whether 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 Fast Reroute (FRR) [RFC4090]) is in place for the link requires careful consideration and coordination of defect detection intervals for the link and the tunnel. When using multi-layer protection, particular consideration must be given to the interaction of defect detections at different network layers. It is recommended to use longer detection intervals at the higher layers. Some recommendations suggest using a multiplier of 3 or larger, e.g., 10 msec detection for the link failure detection and at least 100 msec for the tunnel failure detection. In many cases, it is not practical to use both protection methods simultaneously 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 Label Switched Path (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 to be a valid candidate UMH.

3.1.4. Leaf-Initiated P-Tunnels

An Upstream PE MUST 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 traffic counter information per (C-S,C-G) 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 execute its actions. Determining that a tunnel is probably down by waiting for enough packets to fail to arrive as expected is a heuristic and operational matter that depends on the maximum inter-packet time. A timeout of three seconds is a generally suitable default waiting period to ascertain that the tunnel is down, though other values would be needed for atypical conditions.

In cases where this mechanism is used in conjunction with the method described in Section 5, no prior knowledge of the rate or maximum inter-packet time on the multicast streams is required; downstream PEs can periodically compare actual packet reception statistics on the two P-tunnels to determine when one of them is down. The detailed specification of this mechanism is outside the scope of this document.

3.1.6. BFD Discriminator Attribute

The 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. A P2MP BFD session can be instantiated using a mechanism other than the BFD Discriminator attribute, e.g., MPLS LSP Ping ([MPLS-P2MP-BFD]). The description of these methods is outside the scope of this document.

This document defines the format and ways of using a new BGP attribute called the "BFD Discriminator" (38). It is an optional transitive BGP attribute. Thus, it is expected that an implementation that does not recognize or is configured not to support this attribute, as if the attribute was unrecognized, follows procedures defined for optional transitive path attributes in Section 5 of [RFC4271]. See Section 7.2 for more information. The format of this attribute is shown in Figure 1.

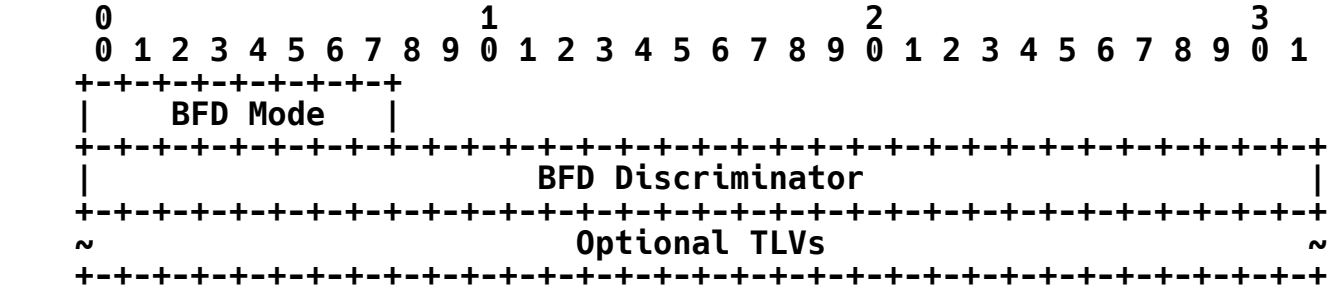

Figure 1: Format of the BFD Discriminator Attribute

Where:

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

BFD Discriminator field is 4 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:

Type: a 1-octet-long field that characterizes the interpretation of the Value field (Section 7.3)

Length: a 1-octet-long field equal to the length of the Value field in octets

Value: a variable-length field

All multibyte fields in TLVs defined in this specification are in network byte order.

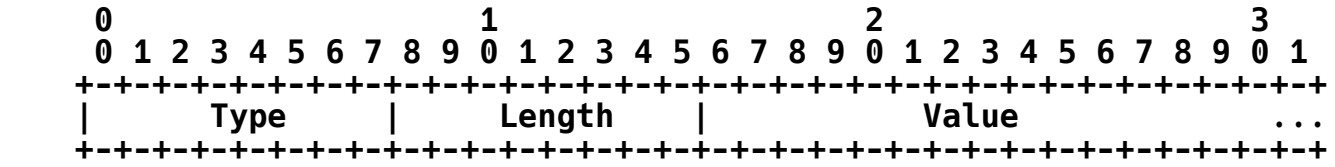

Figure 2: Format of the Optional TLV

An optional Source IP Address TLV is defined in this document. The Source IP Address TLV MUST be used when the value of the BFD Mode field's value is P2MP BFD Session. The BFD Discriminator attribute that does not include the Source IP Address TLV MUST be handled according to the "attribute discard" approach, as defined in [RFC7606]. For the Source IP Address TLV, fields are set as follows:

* The Type field is set to 1 (Section 7.3).

* The Length field is 4 for the IPv4 address family and 16 for the IPv6 address family. The TLV is considered malformed if the field is set to any other value.

* The Value field contains the address associated with the MultipointHead of the P2MP BFD session.

The BFD Discriminator attribute MUST be considered malformed if its length is smaller than 11 octets or if Optional TLVs are present but not well formed. If the attribute is deemed to be 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];

* when transmitting BFD Control packets MUST set the IP destination address of the inner IP header to the internal loopback address 127.0.0.1/32 for IPv4 [RFC1122]. For IPv6, it MUST use the loopback address ::1/128 [RFC4291];

* MUST use the IP address included in the Source IP Address TLV of the BFD Discriminator attribute 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 the 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 that 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 resent with the only change between the previous advertisement and the new advertisement to be the inclusion of the BFD Discriminator attribute.

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:

* the x-PMSI A-D Route MUST be resent with the only change between the previous advertisement and the new advertisement be the exclusion of the BFD Discriminator attribute;

* the P2MP BFD session MUST be deleted. The session MAY be deleted after some configurable delay, which should have a reasonable default.

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];

* to properly demultiplex BFD session, MUST use:

  - the IP address in the Source IP Address TLV included the BFD Discriminator attribute in the x-PMSI A-D Route;

  - the value of the BFD Discriminator field in the BFD Discriminator attribute;

  - the x-PMSI Tunnel Identifier [RFC6514] the BFD Control packet was received on.

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 if the Detection Timer associated with the BFD session 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 to be 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 to be in the Up state. That MAY be determined from the state of a P2MP BFD session with the Standby Upstream PE as the MultipointHead.

If the downstream PE's P-tunnel is already established when the downstream PE receives the new x-PMSI A-D Route with the BFD Discriminator attribute, the downstream PE MUST associate the value of the 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 set up using the BFD Discriminator attribute, and both 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, then:

* The downstream PE MUST stop processing BFD Control packets for this P2MP BFD session;

* The P2MP BFD session associated with the P-tunnel MUST be deleted. The session MAY be deleted after some configurable delay, which should have a reasonable default.

* The downstream PE MUST NOT switch the traffic to the Standby Upstream PE.

3.1.7. BFD Discriminator per PE-CE Link

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 MUST set the bfd.LocalDiag of the P2MP BFD session to Concatenated Path Down or Reverse Concatenated Path Down (per Section 6.8.17 of [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.

3.1.8. Operational Considerations for Monitoring a P-Tunnel's Status

Several methods to monitor the status of a P-tunnel are described in Section 3.1.

Tracking the root of an MVPN (Section 3.1.1) reveals the status of a P-tunnel based on the control plane information. Because, in general, the MPLS data plane is not fate sharing with the control plane, this method might produce false-positive or false-negative alarms, for example, resulting in tunnels that are considered Up but are not able to reach the root, or ones that are declared down prematurely. On the other hand, because BGP next-hop tracking is broadly supported and deployed, this method might be the easiest to deploy.

The method described in Section 3.1.2 monitors the state of the data plane but only for an egress P-PE link of a P-tunnel. As a result, network failures that affect upstream links might not be detected using this method and the MVPN convergence would be determined by the convergence of the BGP control plane.

Using the state change of a P2MP RSVP-TE LSP as the trigger to re-evaluate the status of the P-tunnel (Section 3.1.3) relies on the mechanism used to monitor the state of the P2MP LSP.

The method described in Section 3.1.4 is simple and is safe from causing false alarms, e.g., considering a tunnel operationally Up even though its data path has a defect or, conversely, declaring a tunnel failed when it is unaffected. But the method applies to a

subset of MVPNs, those that use the leaf-triggered x-PMSI tunnels.

Though some MVPNs might be used to provide a multicast service with predictable inter-packet intervals (Section 3.1.5), the number of such cases seem limited.

Monitoring the status of a P-tunnel using a P2MP BFD session (Section 3.1.6) may produce the most accurate and expedient failure notification of all monitoring methods discussed. On the other hand, it requires careful consideration of the additional load of BFD sessions onto network and PE nodes. Operators should consider the rate of BFD Control packets transmitted by root PEs combined with the number of such PEs in the network. In addition, the number of P2MP BFD sessions per PE determines the amount of state information that a PE maintains.

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. In the case where more than two PEs are connected to the C-S site, selection of the Standby PE can be performed using one of the methods of selecting a UMH. Details of the selection are outside the scope of this document. The procedures require all the PEs of that MVPN to follow the same UMH selection procedure, as specified in [RFC6513], regardless of whether the PE selected based on its IP address, the hashing algorithm described in Section 5.1.3 of [RFC6513], or the Installed UMH Route. The consistency of the UMH selection method used among all PEs is expected to be provided by the management plane. 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 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 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, 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. Selection of the behavior, revertive or non-revertive, is an operational issue, but it MUST be consistent on all PEs in the given MVPN. While revertive is considered the default behavior, there might be cases where the switchover to the standby tunnel does not affect other services and provides the required quality of service. In this case, an operator might use non-revertive behavior to avoid unnecessary switchover and thus minimize disruption to the multicast service.

For readability, in the following subsections, 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:

* 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);

* It MUST carry the "Standby PE" BGP Community (0xFFFF0009); see Section 7.1.

The Local Preference attribute of both the normal and the standby C-multicast route needs to be adjusted as follows: 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, 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 the value lower than the value specified as the LOCAL_PREF attribute for the route that does not carry the Standby PE BGP Community. The value of zero is RECOMMENDED.

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 resends the C-multicast route with the 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 absence of the Standby PE BGP Community). The new Upstream PE must set the LOCAL_PREF attribute for that C-multicast route to the same value as when the Standby PE BGP Community was included in the advertisement.

4.2. Upstream PE Behavior

When a PE supporting this specification receives a C-multicast route for a particular (C-S,C-G) for which all of the following are true:

* the RT carried in the route results in importing the route into a particular VRF on the PE;

* the route carries the Standby PE BGP Community; and

* the PE determines (via a method of failure detection that is outside the scope of this document) that C-S is not reachable through some other PE (more details are in Section 4.3),

then the PE MAY 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 MAY 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); and

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 step a.)

Doing neither step a nor step b for a given (C-S,C-G) is called "cold root standby".

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

Doing step b (which implies also doing step 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; that is, it shall not wait for receiving a non-Standby C-multicast route before advertising the corresponding S-PMSI.

Section 9.3.2 of [RFC6513] 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:

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

* 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);

* other mechanisms

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 MVPNs 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.

Prerequisites for the procedures described below to be applied for a source of a given MVPN are:

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

* 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 autoconfigured 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.

In additional 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 of [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 RTs, 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 implementation that supports this specification MUST be configurable to perform as follows:

* 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).

* If the route was received over eBGP and its MED attribute is 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 and, when applied, MAY modify the above-listed behavior.

5. Hot Root Standby

The mechanisms defined in Sections 3 and 4 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 will thus start forwarding traffic for a given multicast state after they have a (primary) BGP C-multicast route or a Standby BGP C-multicast route for that state (or both).

* A policy controls from which tunnel downstream PEs accept traffic. For example, the policy could be based on the status of the tunnel or tunnel-monitoring method (Section 3.1.5).

Other combinations of the mechanisms proposed in Sections 3 and 4 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 Extended

Community of each VPN route from each Upstream PE). 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 MVPN 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 5) and can also be the case if the procedures of Section 3 are used; likewise, it can also be the case when 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 has allocated the BGP "Standby PE" community value 0xFFFF0009 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 has allocated codepoint 38 in the "BGP Path Attributes" registry to the BFD Discriminator attribute.

IANA has created a new "BFD Mode" subregistry in the "Border Gateway Protocol (BGP) Parameters" registry. The registration policies, per [RFC8126], for this subregistry 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" Subregistry Registration Policies

IANA has made initial assignments according to Table 2.

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

Table 2: "BFD Mode" Subregistry

7.3. BFD Discriminator Optional TLV Type

IANA has created a new "BFD Discriminator Optional TLV Type" subregistry in the "Border Gateway Protocol (BGP) Parameters" registry. The registration policies, per [RFC8126], for this subregistry 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 TLV Type" Subregistry Registration Policies

IANA has made initial assignments according to Table 4.

Value	Description	Reference
0	Reserved	This document
1	Source IP Address	This document
2- 175	Unassigned	
176 - 249	Unassigned	

250 - 254	Experimental Use	This document
255	Reserved	This document

Table 4: "BFD Discriminator Optional TLV Type" Subregistry

8. Security Considerations

This document describes procedures based on [RFC6513] and [RFC6514]; hence, it shares the security considerations respectively represented in those 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 provide a mechanism to limit the overall amount of capacity used by the BFD traffic (as the combination of the number of active P2MP BFD sessions and the rate of BFD Control packets to process).

The methods described in Section 3.1 may produce false-negative state changes that can be the trigger for an unnecessary convergence in the control plane, ultimately negatively impacting the multicast service provided by the VPN. An operator is expected to consider the network environment and use available controls of the mechanism used to determine the status of a P-tunnel.

9. References

9.1. 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>.

9.2. Informative References

[MPLS-P2MP-BFD] Mirsky, G., Mishra, G., and D. Eastlake 3rd, "BFD for Multipoint Networks over Point-to-Multi-Point MPLS LSP", Work in Progress, Internet-Draft, draft-mirsky-mpls-p2mp-bfd-14, March 2021, <https://tools.ietf.org/html/draft-mirsky-mpls-p2mp-bfd-14>.

[RFC1122] Braden, R., Ed., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, DOI 10.17487/RFC1122, October 1989, <https://www.rfc-editor.org/info/rfc1122>.

[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>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[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>.

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.

Contributors

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
2018 Antwerp
Belgium

Email: wim.henderickx@nokia.com

Pradeep Jain
Nokia
701 E Middlefield Rd
Mountain View, CA 94043
United States of America

Email: pradeep.jain@nokia.com

Jayant Kotalwar
Nokia
701 E Middlefield Rd
Mountain View, CA 94043
United States of America

Email: Jayant.Kotalwar@nokia.com

Praveen Muley
Nokia
701 East Middlefield Rd
Mountain View, CA 94043
United States of America

Email: praveen.muley@nokia.com

Ray (Lei) Qiu
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
United States of America

Email: rqiu@juniper.net

Yakov Rekhter
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
United States of America

Email: yakov@juniper.net

Kanwar Singh
Nokia
701 E Middlefield Rd
Mountain View, CA 94043
United States of America

Email: kanwar.singh@nokia.com

Authors' Addresses

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

Email: thomas.morin@orange.com

Robert Kebler (editor)
Juniper Networks
1194 North Mathilda Avenue
Sunnyvale, CA 94089
United States of America

Email: rkebler@juniper.net

Greg Mirsky (editor)
ZTE Corp.

Email: gregimirsky@gmail.com