Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 2, 2015

T. Morin, Ed.
Orange
R. Kebler, Ed.
Y. Rekhter
R. Qiu
Juniper Networks
R. Aggarwal
Arktan
W. Henderickx
P. Muley
Alcatel-Lucent

July 1, 2014

Multicast VPN fast upstream failover
draft-morin-l3vpn-mvpn-fast-failover-06

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 advertized toward a standby upstream PE.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 http://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 January 2, 2015.

Table of Contents

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.

The sections 3 and 4 describe two independent mechanisms, allowing different levels of resiliency, and providing different failure coverage:

- 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 [I-D.mofrr] 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].

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 said (C-S,C-G).

The procedure described here is an OPTIONAL procedure that consist in having a downstream PE take into account the status of P-tunnels rooted at each possible upstream PEs, for including or not including each said PE in the list of candidate UMHs for a said (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.

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 said (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 I- or S- PMSI applicable to the said (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 considering invalid some UMH PEs that use a policy where no I-PMSI is advertized for a said 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 advertized 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 said downstream PE and a said VRF, the P-tunnel corresponding to a said upstream PE for a said (C-S,C-G) state is the S-PMSI tunnel advertized 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 advertized by that PE and imported into that VRF.

Note that this documents 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 procedure proposed here also allows 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*.

Depending on the criteria used to determine the status of a P-tunnel, there may be an interaction with other 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 said tunnel happens to be the same as the next-hop address in the BGP autodiscovery 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 said (S,G) if the P-tunnel for this S,G (I or S , depending) 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. P2MP LSP OAM

When a P2MP connectivity verification mechanism such as [I-D.katz-ward-bfd-multipoint] used in conjunction with bootstrapping mechanisms described in [I-D.ietf-mpls-mcast-cv] has been setup for a tunnel, the result of the connectivity verification can be used to define the status of the tree.

If a MultipointHead session has been established on a P2MP MPLS LSP so that BFD packets are periodically sent from the root toward leaves, a condition to consider the status of corresponding tunnel as up is that the BFD SessionState is Up.

When such a procedure is used, in 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.

3.1.6. (S,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 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 (S,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.

4. Standby C-multicast route

The procedures described below are limited to the case where the site that contains C-S is connected to exactly two PEs. The procedures require all the PEs of that MVPN to follow the single forwarder PE selection, as specified in [RFC6513]. 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 said 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.

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 semantic 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 advertized by the standby PE for S)

- 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 a Standby BGP C-multicast route MUST have the LOCAL_PREF attribute set to zero.

Note that, when a PE advertizes such a Standby C-multicast join for an (S,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), nor (b) for a said (C-S,C-G) is called "cold root standby".

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

Doing (b) (which implies also doing (a)) for a said (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 (S,G) as soon as it receives a C-multicast route for (S,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 a 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 said MVPN is:

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

- that these Inter-AS I-PMSI autodiscovery 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 autodiscovery route advertized 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 autodiscovery route advertized 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-advertized 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-advertized 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 said VRF (or possibly only for a said 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 said 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 advertized 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 usptream 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 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)

[Note to RFC Editor: this section may be removed on publication as an RFC.]

8. Security Considerations

9. Acknowledgements

The authors want to thank Greg Reaume and Eric Rosen for their review and useful feedback.

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

[RFC6513] Aggarwal, R., Bandi, S., Cai, Y., Morin, T., Rekhter, Y., Rosen, E., Wijnands, I., and S. Yasukawa, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

10.2. Informative References

[I-D.ietf-mpls-mcast-cv]
Swallow, G., "Connectivity Verification for Multicast Label Switched Paths", draft-ietf-mpls-mcast-cv-00 (work in progress), April 2007.

[I-D.katz-ward-bfd-multipoint]
Katz, D. and D. Ward, "BFD for Multipoint Networks", draft-katz-ward-bfd-multipoint-02 (work in progress), February 2009.

[I-D.mofrr]
Karan, A., Filsfils, C., Farinacci, D., Decraene, B., Leymann, N., and T. Telkamp, "Multicast only Fast Re-Route", draft-ietf-rtgwg-mofrr-04 (work in progress), November 2014.

[RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

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

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

Email: yakov@juniper.net

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

Email: rqiu@juniper.net

Rahul Aggarwal
Arktan

Email: raggarwa_1@yahoo.com

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
Antwerp 2018
Belgium

Email: wim.henderickx@alcatel-lucent.com

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

Email: praveen.muley@alcatel-lucent.com