BESS
Internet-Draft
Updates: 6513, 6514 (if approved)
Intended status: Standards Track
Expires: 8 April 2023

Z. Zhang
R. Kebler
W. Lin
Juniper Networks
E. Rosen

5 October 2022

MVPN/EVPN C-Multicast Routes Enhancements
draft-zzhang-bess-mvpn-evpn-cmcast-enhancements-02

Abstract

[RFC6513] and [RFC6514] specify procedures for originating, propagating, and processing "C-multicast routes". However, there are a number of MVPN use cases that are not properly or optimally handled by those procedures. This document describes those use cases, and specifies the additional procedures needed to handle them. Some of the additional procedures are also applicable to EVPN SMET routes [RFC9251].

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 [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 https://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on 8 April 2023.

Copyright Notice

Table of Contents

1. Introduction

[RFC6513] and [RFC6514] specify procedures for originating, propagating, and processing "C-multicast routes". However, there are a number of MVPN use cases that are not properly or optimally handled by those procedures. This document describes those use cases, and specifies the additional procedures needed to handle them.

Some of the additional procedures are also applicable to EVPN SMET routes [RFC9251]; this is discussed in Section 1.5.

1.1. Terminology

This document uses terminology from MVPN and EVPN. It is expected that the audience is familiar with the concepts and procedures defined in [RFC6513], [RFC6514], [RFC7524], [RFC7432], [I-D.ietf-bess-evpn-bum-procedure-updates], and [RFC9251]. Some terms are listed below for references.

* PMSI: P-Multicast Service Interface - a conceptual interface for a PE to send customer multicast traffic to all or some PEs in the same VPN.

* I-PMSI: Inclusive PMSI - to all PEs in the same VPN.

* S-PMSI: Selective PMSI - to some of the PEs in the same VPN.

* C-G-BIDIR: A bidirectional multicast group address (i.e., a group address whose IP multicast distribution tree is built by BIDIR-PIM) in customer address space.

* RBR: Regional Border Router. A provider tunnel could be segmented, with one segment in each region. A region could be an AS, an IGP area, or even a subarea.

1.2. MVPN C-Bidir Support with VPN Backbone being RPL

In BIDIR-PIM [RFC5015], every group is associated with a "Rendezvous Point Link" (RPL). The RPL for a given group G is at the root of the BIDIR-PIM distribution tree. Links of the distribution tree that lead towards the RPL are considered to be "upstream" links, and links that lead away from the RPL are considered to be "downstream" links. Every node on the distribution tree has one upstream link and zero or

more downstream links.

Data addressed to a BIDIR-PIM group may enter the distribution tree at any node. The entry node sends the data on the upstream links and the downstream links. A node that receives the data from a downstream link sends it on its upstream link and on its other downstream links. A node that receives the data from its upstream link sends it on its downstream links. When a node that is attached to the RPL receives data from a downstream link, it forwards the data onto the RPL (as well as onto any other downstream links.) When node attached to the RPL receives data from the RPL, it forwards the data downstream.

The above is a simplified description, and ignores the fact that every link except the RPL has a Designated Forwarder (DF). Only the DF forwards traffic onto the link. However, the RPL has no DF; any node can forward traffic onto the RPL.

1.2.1. C-multicast Routes for the MVPN-RPL Method of C-BIDIR support

Section 11.1 of [RFC6513] describes a method of providing MVPN support for customers that use BIDIR-PIM. This is known as "MVPN C-BIDIR support". In this method of C-BIDIR support, the VPN backbone itself functions as the RPL. Thus this method is known as the "MVPN-RPL" method. The RPL is actually an I-PMSI or S-PMSI. The PE routers treat the I-PMSI or S-PMSI as their upstream link, and treat their VRF interfaces as downstream links.

If the MVPN-RPL method of C-BIDIR support is being used in a particular MVPN, all the PEs attached to that MVPN must be provisioned to use this method.

In the context of a given VPN, a PE with interest in receiving a particular C-BIDIR group (call it C-G-BIDIR) advertises this interest to the other PEs by originating a C-multicast Shared Tree Join route. When any PE receives traffic for the C-G-BIDIR on its PE-CE interface, it sends the data to the MVPN-RPL if and only if it has received corresponding (C-*,C-G-BIDIR) C-multicast Shared Tree Join route. Other PEs receive the traffic on the MVPN-RPL and forward to their downstream receivers. However, the procedure for constructing the C-multicast Shared Tree Join route in this case is not fully specified in [RFC6513] or [RFC6514]. The proper set of procedures are specified in Section 2.1.1 of this document.

Compared to other C-Multicast routes specified in [RFC6514], these are "untargeted" in that the RT allows all PEs in the same MVPN to import them, while those other C-Multicast routes use a RT that identifies a VRF on a particular Upstream Multicast Hop (UMH) PE.

If a PE wants to use selective tunnel to send traffic to only a subset of the PEs on MVPN-RPL, i.e., those with downstream (C-*,C-G-BIDIR) state, per [RFC6513] [RFC6514] the PE needs to advertise a corresponding (C-*,C-G-BIDIR) S-PMSI A-D route, whose PTA specifies the tunnel to be used. In case of RSVP-TE P2MP, Ingress Replication (IR), or BIER tunnel, the Leaf Information Required (LIR) bit in the S-PMSI route's PTA is set to solicit corresponding Leaf A-D routes from those PEs with downstream (C-*,C-G-BIDIR) state. Every PE that wants to use selective tunnel for the (C-*,C-G-BIDIR) will advertise its own S-PMSI A-D route, each triggering a set of corresponding Leaf A-D routes.

Notice that the (C-*,C-G-BIDIR) C-Multicast routes from different PEs all have their own RDs so Route Reflectors (RRs) will reflect every one of them, and they already serve explicit tracking purpose (the BGP Next Hop identifies the originator of the route in non-segmentation case) - there is no need to use Leaf A-D routes triggered by the LIR bit in S-PMSI A-D routes. In case of RSVP-TE P2MP tunnel, the S-PMSI A-D routes are still needed to announce the tunnel but the LIR bit does not need to be set. In case of IR/BIER, there is no need for S-PMSI A-D routes at all.

1.2.2. Optional use of MVPN-RPL RD with mLDP/PIM Provider Tunnels

When mLDP/PIM tunnels are used, there is no need for explicit tracking as the leaves will simply send mLDP label Mapping or PIM Join messages. As a result, it's unnecessary for a PE to retain each C-Multicast route from each PE for the same C-G-BIDIR. If there is a Route Reflector (RR) in use, and it is known apriori that all the PEs/RRs/ASBRs involved in the propagation of the C-Multicast routes support BGP ADD-PATH [RFC7911], then the PEs could use a common RD to construct the C-Multicast routes. That way, the routes from different PEs for the same C-G-BIDIR will be considered paths for the same route and the RRs will reflect N paths to each PE. If N is significantly smaller than the number of PEs that advertises the routes, then the burden is significantly reduced for the PEs.

The reason for the need for ADD-PATH is shown with this example: both PE1 and PE2 advertise the same (C-*,C-G-BIDIR) C-Multicast route and the RR chooses the one from PE1 as the active path. Without ADD-PATH, the RR won't reflect any (C-*,C-G-BIDIR) path back to PE1, causing PE1 to think there is no other PE interested in receiving the C-G-BIDIR traffic. With ADD-PATH, it is guaranteed that even the originator of the active path will receive at least one other path. For this reason, ADD-PATH is needed and N=2 is well enough.

1.2.3. MVPN C-ASM Support without CE Routers

Current MVPN specifications is based on the fact that CEs are routers and in case of ASM one or more of the routers in customer address space, which could be a CE, a PE's VRF, or another non-PE/CE router, serves as RP. Traffic may be delivered on shared trees, switch to source specific trees, or switch back to shared trees depending the situation. There are two modes of MVPN to support ASM, all involving (C-S,C-G) MVPN Source Active (SA) A-D routes, individual (C-S,C-G) control/forwarding plane state and procedures that are not needed for a special scenario where CEs are not routers but just hosts.

From a logical point of view, this special scenario is when a VPN only involves customer networks directly connected to the PEs and no customer routers are used.. A practical example is EVPN inter-subnet multicast [I-D.ietf-bess-evpn-irb-mcast], when EVPN is used to connect only servers and no customer routers are involved. In this case, it does not make sense to introduce the RP concept into the deployment and involve the MVPN SA procedures. Rather, this could be modeled as C-Bidir with MVPN-RPL and all the above discussed optimizations apply.

1.3. Inter-AS Propagation of MVPN C-Multicast Routes

Section 11.2 of [RFC6514] specifies the procedure used to propagate C-multicast routes from one AS to another. However, there are a number of problems with the procedures as specified in that RFC.

RFC6514 presumes that C-multicast routes are propagated through the ASBRs. This is analogous to RFC 4364's "Inter-AS option b". However, in some deployment scenarios, the C-multicast routes are propagated through Route Reflectors, in a manner analogous to RFC 4364's "Inter-AS option c". Strictly speaking, RFC 6514 does not allow this deployment scenario. This document updates RFC 6514 by allowing this deployment scenario to be used in place of the procedures of Section 11.2 of RFC 6514.

In some deployment scenarios, the propagation of C-multicast routes is controlled by the "Route Target Constraint" procedures of [RFC4684]. Strictly speaking, RFC 6514 does not allow this deployment scenario. This document updates RFC 6514 by allowing this deployment scenario to be used in place of the procedures of Section 11.2 of RFC 6514.

Per [RFC6514], an MVPN C-Multicast route is targeted at a particular PE, and its inter-as propagation towards the PE follows a series of ASBRs (in the reverse order) on the propagation path of one of the following:

* The Intra-AS I-PMSI A-D route from the targeted PE, if the deployment is using non-segmented tunnels. In this scenario, the IP address of the targeted PE is encoded into the four-octet "Source AS" field (!) of the C-multicast route's NLRI.

* The Inter-AS I-PMSI A-D route for the AS that the targeted PE is in, if the deployment is using segmented tunnel. In this scenario, the AS number of the source PE is encoded into the "Source AS" field of the C-multicast route's NLRI.

In both cases, the corresponding I-PMSI A-D route is found by looking for an I-PMSI A-D route whose NLRI consists of the C-multicast route's RD prepended to the contents of the C-multicast route's "Source AS" field. If neither Inter-AS nor Intra-AS I-PMSI A-D route is used, e.g. (C-*,C-*) S-PMSI A-D route is used, then the specified procedure will not work.

It must be noted that the RFC 6514 Section 11.2 propagation procedures cannot be applied to untargeted C-multicast routes, and cannot be applied even to targeted C-multicast routes if the infrastructure is based on IPv6 rather than IPv4.

This document updates RFC 6514 by declaring that the procedure of Section 11.2 of that document is only applicable in the case that (1) the C-multicast routes are being propagated through the ASBRs, AND (2) the propagation of those routes is not under the control of the Route Target Constraint procedures. It also updates the procedures of Section 11.2 of [RFC6514] to allow it to work without relying on I-PMSI A-D routes, whether IPv4 or IPv6 infrastructure is used.

This document also updates RFC 6514 by declaring that C-multicast routes MAY be propagated using ordinary BGP propagation procedures, which do not rely on the presence of I-PMSI A-D routes. For targeted C-multicast routes, this will result in a less optimal propagation path, but it does work in all cases. The Route Target Constraint procedures can always be used to obtain a more optimal path.

The selection of the propagation procedure for C-multicast routes is determined by provisioning.

In Section 1.2.1, the explicit tracking using C-multicast route relies on that the route's next hop is not changed so that the next hop can identify the originator. If the c-multicast routes are propagated through ASBRs, the next hop will be changed. With tunnel segmentation, this is not a problem (see Section 1.6) but if non-segmented tunnels are used, either the C-multicast route propagation must follow the Option C procedures and the next hop is not changed by the RRs, or the routes must carry an EC to identify the

originator. Or, the RD of a C-multicast route can be used to locate an I/S-PMSI route from the same PE, in which the Originator IP Address can be found.

1.4. MVPN Inter-AS Upstream PE Selection

Consider the following scenario:

A multicast source is multi-homed to PE1 and PE2 in the same source AS1. ASBR1 in AS1 connects to ASBR2 in another AS2. In AS2, egress PE3 selects PE1 while egress PE4 selects PE2 as their upstream PE respectively, because they use the "Installed UMH Route" as the "Selected UMH Route" (as defined in Section 5.1.3 of [RFC6513]).

Suppose inter-as tunnel segmentation is used. Following Section 11.1.3 of [RFC6514], PE3 and PE4 will construct their C-multicast routes with the same NLRI key (in particular with the same RD from the Inter-AS I-PMSI A-D route originated by ASBR1) but with one different Route Target - PE3's C-multicast route carries the RT corresponding to PE1's VRF while PE4's C-multicast route carries the RT corresponding to PE2's VRF. ASBR2 will re-advertise only one of the two C-multicast routes to ASBR1. Assuming it is the one with a RT corresponding to PE1, then only PE1 will transmit corresponding traffic.

If selective tunnels are used, PE4 that chooses PE2 as the upstream PE will not join the selective tunnel advertised by PE1 so it will not receive traffic.

With the new method for inter-as propagation of C-multicast routes described in the previous section, this traffic blackholing problem can be resolved if PE3 and PE4 construct their C-multicast routes with different RDs, e.g. with the RD from the chosen UMH route instead of the RD from the Inter-AS I-PMSI A-D route. That way, PE1 will receive the C-multicast route from PE3 and PE2 will receive the C-multicast route from PE4. Both will transmit traffic but PE3 and PE4 will only receive the traffic via the selective tunnel that they join hence no duplication or blackholing.

Notice that this also removes the pre-requisite in Section 4.4 of [RFC9026].

However, there are still two problems. First, while there is no duplication or blackholing issue when selective tunnels are used, two copies of traffic are sent inter-AS, possibly through many common paths before reaching the egress PEs (imagine that there are a string of ASes between AS1 and AS2). This is not an efficient use of inter-AS resources.

Choosing upstream PE based on installed UMH route allows different egress PEs to choose different upstream PEs (typically the closest upstream PE), so it is desired for certain intra-as deployment scenarios but apparently it is not desired for PEs in other ASes to choose different upstream PEs. This problem can actually be solved if PEs always do "Single Forwarder Selection" (the default method described in Section 5.1.3 of [RFC6513]) for sources in other ASes while (if provisioned so) selecting upstream PE based on installed UMH routes for sources in the local AS.

The second problem is that, when inclusive inter-as tunnels are used, if both PE1 and PE2 send the same traffic, ASBR1 will inject duplicate traffic into the same inter-as tunnel, while PE3 and PE4 has no way to distinguish the source PE of each copy.

There are two solutions to the second problem. The first solution is that ASBRs advertise PE Distinguisher (PED) labels (Section 8 of [RFC6514]) via a PED attribute attached to their Inter-AS I-PMSI A-D routes, and push a label that identifies the ingress PE when it sends a packet into the inclusive inter-AS tunnel, and an egress PE discards traffic not from its chosen upstream PE.

The other solution is for the ingress ASBR to only accept traffic from one ingress PE and forward into the inclusive inter-as tunnel. This does not require egress PEs to discard traffic based on an additional PED label, but does require the ingress ASBR to participate upstream PE selection and do IP forwarding in a VRF for the source VPN, so that it can choose the copy to accept and forward. Because it may not have local receivers, it needs to receive C-multicast routes from egress PEs who will receive corresponding traffic from it, and import the routes into its local VRF.

1.5. EVPN Selective Multicast Ethernet Tag (SMET) Routes

[RFC9251] defines a new EVPN route type known as an "SMET route".

The EVPN SMET routes are analogous to the MVPN C-muilticast routes, in that both type of routes are used to disseminate the information that a particular egress PE has interest in a particular multicast C-flow or set of C-flows.

An EVPN SMET route contains, in its NLRI, the RD associated with the VRF from which the SMET route was originated. In addition, it is disseminated to all PEs of a given EVI. In this way, SMET routes are analogous to the MVPN C-multicast routes that are used for C-BIDIR support.

An EVPN SMET route contains, in its NLRI, the IP address of the originating PE. In this way, they are analogous to the MVPN Leaf A-D routes (They really combine the function of the MVPN C-multicast routes and the MVPN Leaf A-D routes). Similarly, they are also analogous to the C-multicast route for MVPN-RPL that carries an EC that identifies the originating PE.

In EVPN, as in MVPN, explicit tracking is required when selective tunnels are realized using IR, BIER, or RSVP-TE P2MP. The EVPN SMET routes provide this explicit tracking, so in these cases EVPN does not need explicit Leaf A-D routes. With IR/BIER, there is no need for S-PMSI route either. However, when SMET routes are used with segmented IR/BIER tunnels, more procedures are needed, just like the C-multicast route in MVPN-RPL case (Section 1.6). For that reason, given the similarity between SMET and C-Multicast routes, in this document we will use the same term C-Multicast route for EVPN SMET route as well. The two may be used interchangeably in case of EVPN.

If selective tunnels are set up using procedures that do not require explicit tracking, e.g. mLDP or PIM, the following optimization could be done, similar to MVPN-RPL with mLDP/PIM tunnels (Section 1.2.2):

* When constructing an SMET route, put 0 as the Originator Router Address.

* When constructing an SMET route in the context of a given EVI, have all PEs of that EVI set the RD field of the NLRI to the same value (This is analogous to "MVPN-RPL RD" discussed in Section 1.2.2).

* When a Route Reflector distributes the SMET routes, it uses BGP ADD-PATH to distribute at least two "paths" for a given NLRI.

1.6. Provider Tunnel Segmentation with Explicit-Tracking C-Multicast Routes

For the above MVPN-RPL and EVPN cases where C-multicast routes are used for explicit tracking without requiring corresponding S-PMSI A-D routes in case of IR/BIER selective tunnel, it works well when there is no tunnel segmentation. With tunnel segmentation [RFC6514] [RFC7524], [I-D.ietf-bess-evpn-bum-procedure-updates] additional procedures are needed.

1.6.1. Conventional Tunnel Segmentation

Multicast forwarding needs to follow a rooted tree. With segmentation, the tree is divided into segments, with each segment rooted at either the ingress PE or a Regional Border Router (RBR). A segment is contained in a region, which could be an AS, an area, or a sub-area. The root of a segment only needs to track the leaves in its region, which are PEs or RBRs in that region. With the traditional PMSI/Leaf A-D procedures, an ingress PE/RBR sends out an I/S-PMSI route, propagated by RBRs (segmentation points), who change the tunnel identifier along the way to identify the tunnels for their segments. The Leaf A-D routes from PEs are not propagated by the RBRs. Rather, a RBR will proxy the Leaf AD routes it receives from its downstream towards its upstream RBR or PE, following the I/S-PMSI A-D routes received in the upstream region, as specified in [RFC6514] [RFC7524] [I-D.ietf-bess-evpn-bum-procedure-updates].

1.6.2. Selective Tunnel Segmentation with Untargeted Explicit-Tracking C-multicast Routes

Without segmentation, the untargeted explicit-tracking C-Multicast routes are sent to every PE, and each PE adds the originator of the routes as leaves of the tunnel rooted at the PE.

With segmentation, untargeted explicit-tracking C-Multicast routes are propagated through segmentation points towards all ingress PEs or ASes and are merged along the way. This is like the traditional PMSI/Leaf A-D procedures but with one difference.

With the traditional PMSI/Leaf A-D procedures, the propagation is towards the originator of the PMSI A-D route and a single tree is formed. With untargeted C-Multicast routes, multiple trees are formed, each being rooted at the ingress PE (if per-region aggregation [I-D.ietf-bess-evpn-bum-procedure-updates] is not used) or ingress RBR (if per-region aggregation is used). The roots of those trees are either the ingress PEs or the ingress RBRs, identified by all the per-PE or per-region I-PMSI A-D routes.

To form those multiple trees without requiring S-PMSI A-D routes from the ingress PEs/RBRs, this document proposes that the RBRs convert a C-multicast route originated in its own region to Leaf A-D routes, as if corresponding S-PMSI A-D routes had been received from ingress PEs/RBRs. The details are provided in Section 2.2.

2. Specifications

This section provides detailed specifications for the optional enhancements introduced above.

2.1. MVPN C-Bidir Support with VPN Backbone being RPL

2.1.1. Constructing C-Multicast Share Tree Join route

In the context of a particular VRF, a PE with downstream state for the group C-G-BIDIR originates a C-multicast Shared Tree Join route, referred to as "MVPN-RPL C-multicast Join", when the MVPN-RPL method of C-BIDIR support is being used.

The fields of the route are set as follows:

* RD: See Section 2.1.1.2.

* Source AS: set to zero.

* Multicast Source Length: 4 or 16.

* Multicast Source: set to RPA.

* Multicast Group Length: 4 or 16.

* Multicast Group: BIDIR-PIM group address.

Note that the RD field, and the Route Targets that are attached to the C-multicast route are different than what is specified in [RFC6514]. See following two sections.

2.1.1.1. Setting the Route Targets

Per [RFC6514], when a PE originates a C-multicast route, it "targets" the route to a specific one of the other PEs attached to the same VPN. The IP address of the targeted PE is encoded into a Route Target and attached to the C-mulitcast route. This ensures that the C-multicast route is processed only by the PE to which it is targeted.

However, C-multicast routes used by the MVPN-RPL method are not targeted. Rather, they must be processed by all the other PEs attached to the same MVPN. Thus we refer to these routes as "untargeted". The Route Targets attached to these routes must be such as to cause the routes to be propagated to all the other PEs of the given MVPN. By default, these will be the same Route Targets that are attached to the I-PMSI A-D routes of the MVPN.

2.1.1.2. Setting the Route Distinguisher

Per [RFC6514], the RD in a C-multicast Join Route is the RD of a VRF on the PE to which the route is targeted. However, in an MVPN-RPL C-multicast Join, the RD is set differently.

If PIM/mLDP provider tunnels are used, and it is known that all the PEs/RRs/ASBRs involved in the propagation of C-multicast routes support BGP ADD-PATH, the RD MAY be set to a value that is specially configured to be used as the RD for MVPN-RPL in a given VPN. Call this the "MVPN-RPL" RD for that VPN. In that case, all the C-multicast Joins that are providing C-BIDIR support (for a given VPN) using the MVPN-RPL method will have the same RD. This MVPN-RPL RD of a given VPN MUST NOT be used for any other purpose, or by any other VPN. See Section 1.2.2 for a discussion of when it may be advantageous to use an MVPN-RPL RD.

For other provider tunnel types, or if the above mentioned MVPN-RPL RD in case of PIM/mLDP tunnel is not feasible (e.g. BGP ADD-PATH is not supported), the RD in the C-multicast route is that of the VRF from which the route is originated.

For Global Table Multicast (GTM) using MVPN procedures [RFC7716], RFC 7716 specifies that MVPN routes use a special 0:0 RD. This document specifies that GTM use non-0:0 RDs for C-Multicast routes for C-Bidir, when the backbone is used as RPL and provider tunnels are not set up by PIM/mLDP.

2.1.2. Setting Up the MVPN-RPL

By default, the I-PMSI or (C-*,C-BIDIR) S-PMSI plays the role of MVPN-RPL. When (C-*,C-G-BIDIR) S-PMSI is used for a particular C-G-BIDIR, the following procedures are followed, depending on the type of provider tunnel used.

2.1.2.1. Ingress Replication or BIER

If Ingress Replication or BIER is used, there is no need for the ingress PE to advertise (C-*,C-G-BIDIR) S-PMSI A-D route. The ingress PE identifies the tunnel leaves to send traffic to by the C-multicast routes it receives, because each such route has a different RD and serves explicit tracking purpose. In case of IR, the label in the Intra-AS I-PMSI A-D route or (C-*,C-*) S-PMSI A-D route from a leaf is used to send traffic to the leaf. In case of BIER, the label in the same route from the ingress PE is used to send traffic.

2.1.2.2. RSVP-TE P2MP

With RSVP-TE P2MP tunnel, the ingress PE advertises (C-*,C-G-BIDIR) S-PMSI A-D route without setting the LIR bit in the route's PTA. It identifies the tunnel leaves from the C-multicast routes it receives.

2.1.2.3. PIM/mLDP

With PIM or mLDP P2MP provider tunnel, procedures in [RFC6514] are followed.

2.2. Inter-AS Propagation of MVPN C-Multicast Routes

This specification allows two methods of Inter-AS propagation for MVPN C-multicast routes. The choice of which method is used is by provisioning.

2.2.1. Procedures in Section 11.2 of [RFC6514]

The procedures in Section 11.2 of [RFC6514] are extended with the following.

The Source AS field in the NLRI of C-multicast route is set to the AS number of the UMH PE if and only if segmented inter-AS tunnels and per-AS aggregation (via Inter-AS I-PMSI A-D routes) are used. The existing procedures are used as is in this case.

Otherwise, when an egress PE constructs a C-Multicast route and the upstream PE is in a different AS from the local PE, it finds in its VRF an Intra-AS I-PMSI A-D route or any S-PMSI A-D route from the upstream PE (the Originating Router's IP Address field of that route has the same value as the one carried in the VRF Route Import of the (unicast) route to the address carried in the Multicast Source field). The RD of the found I/S-PMSI A-D route is used as the RD of the advertised C-multicast route. The Source AS field in the C-multicast route is set to 0. If the Next Hop of the found I/S-PMSI A-D route is an EBGP neighbor of the local PE, then the PE advertises the C- multicast route to that neighbor. Otherwise the PE advertises the C-multicast route into IBGP.

When an ASBR receives a C-multicast route with the Source AS field set to 0, it uses the RD of the C-multicast route to locate an Intra-AS I-PMSI A-D route or any S-PMSI A-D route, and propagate the C-multicast route to the bgp neighbor from which the found I/S-PMSI A-D route is learned.

2.2.2. Ordinary BGP Propagation Procedures

This document specifies that C-multicast routes MAY be propagated using ordinary BGP propagation procedures, which do not rely on the presence of any I/S-PMSI A-D routes. With this method, the construction of C-Multicast A-D routes always follows the same procedures, whether the source is in the same or different AS. Specifically, the 3rd and 5th paragraphs of Section 11.1.3 of [RFC6514] are quoted here:

From the selected UMH route, the local PE extracts (a) the ASN of the upstream PE (as carried in the Source AS Extended Community of the route), and (b) the C-multicast Import RT of the VRF on the upstream PE (the value of this C-multicast Import RT is the value of the VRF Route Import Extended Community carried by the route). The Source AS field in the C-multicast route is set to that AS. The Route Target Extended Community of the C-multicast route is set to that C-multicast Import RT.

...

... the RD of
the advertised MCAST-VPN NLRI is set to the RD of the VPN-IP route that contains the address carried in the Multicast Source field.

For targeted C-multicast routes, this will result in a less optimal propagation path, but it does work in all cases. The Route Target Constraint procedures can always be used to obtain a more optimal path.

2.3. Inter-AS Upstream PE Selection

This document allows that, when selecting upstream PE for a source not in the local AS, the Single Forwarder Selection method, i.e., the default procedure in Section 5.1.3 of [RFC6513] is used, even if the method of using the installed UMH route as the selected UMH route is provisioned (to be used for sources in the local AS only).

2.4. Duplication Prevention on the Same Inclusive Inter-AS Tunnel

The procedures in this section are only applicable when inclusive inter-AS tunnels advertised in Inter-AS I-PMSI A-D routes are used and it is known that an ingress ASBR may receive duplicate traffic from different ingress PEs in the same local AS. One of the following two methods is provisioned consistently on all PEs and ingress ASBRs of a VPN.

2.4.1. Using PE Distinguisher Labels

With this method, an ingress ASBR that may receive duplicate traffic from different PEs and inject into the same inclusive inter-AS tunnels use a PED label to identify the upstream PE of the traffic, so that egress PEs can discard traffic not from their selected upstream PE.

When an ASBR advertises an Inter-AS I-PMSI A-D route, it includes a PE Distinguisher (PED) Labels attribute [RFC6514]. The attribute lists one label for each PE in the corresponding AS, and the labels are allocated from a Domain-wide Common Block (DCB, [I-D.ietf-bess-mvpn-evpn-aggregation-label]). When an ingress ASBR forwards traffic it receives from a local ingress PE, it needs to push the label assigned to the ingress PE and advertised in the PED attribute of corresponding Inter-AS I-PMSI A-D route. Because the labels are assigned from the DCB, they do not need to be swapped along the way. Downstream and upstream assigned labels could be used as well, but that requires the ASBRs swap PED labels along the way (in addition to tunnel label swapping) so they are not discussed here.

Note that if intra-AS tunnel aggregation is used in the ingress AS, the ingress PE SHOULD use the same PED label and the ingress ASBR MUST NOT push the PED label again when forwarding traffic into the inclusive inter-as tunnel.

2.4.2. Ingress ASBR Filtering Out Duplications

With this method, an ingress ASBR performs IP forwarding for traffic that goes onto inclusive tunnels [I-D.zzhang-bess-mvpn-evpn-segmented-forwarding] after discarding traffic not from the upstream PE that it chooses.

The ingress ASBR MUST be provisioned with a VRF for each VPN with local PEs, and with a C-multicast Import RT for the VRF. The Inter-AS I-PMSI A-D route that it advertises for the VPN MUST carry a VRF Route Import Extended Community (EC) that has the value of the C-multicast Import RT for the VRF. This is similar to that a PE includes a VRF Route Import EC in VPN-IP routes that it originates.

When an egress PE constructs a C-multicast routes, if the source is in a different AS, the ingress ASBR that advertises the Inter-AS I-PMSI A-D rotue installed by this egress PE is chosen as the upstream PE. The RD and AS number in the Inter-AS I-PMSI A-D route are used to construct the C-multicast route, and a C-multicast Import RT (for importing the constructed C-multicast route into the ingress ASBR's VRF) is included, with the value of this RT being the value of the VRF Route Import EC carried by the Inter-AS I-PMSI A-D route.

When an ingress ASBR receives a C-multicast route and imports the route into one of its local VRFs (because of the RT constructed as above), it treats as if a PIM/IGMP join was received on the inter-AS inclusive tunnel. It selects its own upstream PE and originates a corresponding C-multicast route. Corresponding traffic received from the selected upstream PE is then routed into the inter-AS inclusive tunnel.

2.5. Provider Tunnel Segmentation with Explicit-Tracking C-Multicast Routes

This section applies when IR/BIER are used for MVPN/EVPN selective tunnels.

If per-region aggregation [I-D.ietf-bess-evpn-bum-procedure-updates] is used, this document specifies that the per-region I-PMSI A-D route MUST carry a VRF Route Import EC to identify the originator of the per-region I-PMSI A-D route. Note that, while it borrows "VRF Route Import EC" from the UMH routes, it is only used to identify the originator.

If per-region aggregation is not used, this document specifies that either per-PE I-PMSI or (C-*,C-*) S-PMSI A-D routes MUST be originated by every PE.

2.5.1. Egress PEs and RBRs

An egress PE originates MVPN C-multicast routes for MVPN-RPL as specified in previous sections of this document, or EVPN SMET routes as specified in [RFC9251]. Recall that EVPN SMET routes may also be referred to C-Multicast routes in this document.

Explicit-tracking C-multicast routes must be processed by segmentation points, which are referred to as RBRs. When a RBR receives a C-multicast route from within its own region, and the route does not carry a flag bit that indicates the route is converted from a downstream Leaf A-D route (see descriptions below), it converts the C-multicat route into one or more Leaf A-D routes, as if it had received corresponding S-PMSI A-D routes. When a converted Leaf A-D routes reaches the ingress region, the RBR converts it back to C-multicast routes.

With per-region aggregation, the RBR in an egress region finds all active per-region I-PMSI A-D route that the RBR has in the corresponding VRF. For each of them, it makes up a (C-S,C-G) or (C-*,C-G) S-PMSI A-D route as following.

* RD: set to the RD from the per-region I-PMSI A-D route.

* Source/Group length/address fields: set according to the received C-multicast route.

* Originator's IP Address: set according to the VRF Route Import EC in the per-region I-PMSI A-D route

* Ethernet Tag ID in case of EVPN: set according to the received SMET route (which is also referred to as C-multicast route).

* Next Hop: set according to the per-region I-PMSI A-D route.

Without per-region aggregation, a RBR finds all active per-PE I-PMSI or (C-*,C-*) S-PMSI A-D route in the VRF. For each of them it makes up a (C-S,C-G) or (C-*,C-G) S-PMSI A-D route similar to the per-region aggregation case. The only difference is that the Originator's IP Address field is set to the same as in the per-PE I-PMSI or (C-*,C-*) S-PMSI A-D route.

A corresponding Leaf A-D route is then generated and propagated to the upstream identified by the BGP next hop in the made up S-PMSI A-D route, following existing PMSI/Leaf A-D route procedures.

If the egress region uses Ingress Replication, the made up S-PMSI A-D route is not propagated anywhere. If the egress region uses PIM or RSVP-TE/mLDP P2MP tunnel, the S-PMSI A-D route is advertised into the egress region to announce the tunnel to be used. If the egress region uses BIER or aggregated RSVP-TE/mLDP P2MP tunnel, the S-PMSI A-D route is also advertised into the egress region and carry an upstream allocated label. The label may be at the per S-PMSI A-D route level or at per VPN/BD level. In the former case, label switching at the RBR can be used. In the latter case, IP lookup in the corresponding VRF or BD is needed.

2.5.2. Transit RBRs

When an upstream RBR receives a (C-S,C-G) or (C-*,C-G) Leaf A-D route, It locates the active per-PE/region I-PMSI or (C-*,C-*) S-PMSI A-D route whose RD matches the received Leaf A-D route. If no such route exists, the received Leaf A-D route is ignored until such a route appears later. It also tries to locate a corresponding active (C-S,C-G) or (C-*,C-G) S-PMSI A-D route, which could be a real one received from an upstream PE/RBR, or could be a made up one triggered by a Leaf A-D route from a different downstream. If such route exists, existing PMSI/Leaf A-D route procedures are followed.

If no such corresponding active (C-S,C-G) or (C-*,C-G) S-PMSI A-D route exists, and the located active I-PMSI or (C-*,C-*) S-PMSI A-D route has a next hop different from the Originator IP Address in the per-PE I-PMSI A-D route or (C-*,C-*) I-PMSI A-D route, or different from the address in the VRF Route Import EC in the per-region I-PMSI A-D route, the ingress region corresponding to the I-PMSI or (C-*,C-*) S-PMSI A-D route has not been reached. The RBR then makes up a (C-S,C-G) or (C-*,C-G) S-PMSI A-D route. as specified earlier, and proxies Leaf A-D routes further up. Similarly, the S-PMSI A-D route may be advertised into the transit region.

2.5.3. Ingress RBRs

If the BGP next hop in the located active I-PMSI or (C-*,C-*) S-PMSI A-D route matches the Originator IP Address in the per-PE I/S-PMSI A-D route or the IP address in the per-region I-PMSI A-D route's VRF Route Import EC, it means the ingress region has been reached. If the corresponding (C-S,C-G) or (C-*,C-G) S-PMSI A-D route is a made up one and not actually advertised by an ingress PE/RBR, and the RBR does not have corresponding local (C-S,C-G) or (C-*,C-G) state, it reconverts the Leaf A-D route back to C-multicast route, with a CV ("Converted") flag bit indicating that the route is not from local state learned on PE-CE interface but from state learned further downstream. The flag bit prevents other RBRs in this region to trigger Leaf A-D routes from this converted C-multicast route.

The converted C-multicast route is constructed as following:

* RD: set to the RD of the RBR for the related IP/MAC VRF.

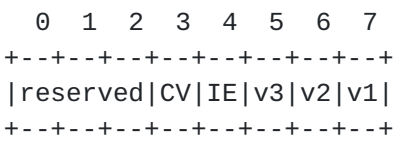
* Source/Group length/address fields: set according to the received Leaf A-D route.

* Ethernet Tag ID in case of EVPN: set according to the received Leaf A-D route.

* Next Hop: set to the RBR's local IP Address.

The RT of the converted C-multicast route is set to the RT used for VRF but the route is only propagated to PEs/RBRs in the local region.

For EVPN SMET routes, the flag bit is part of the existing Flags field in the NLRI:



The IE/v3/v2/v1 are existing bits and the CV bit is the new bit to indicate that this is converted from state learned from downstream.

For MVPN C-Multicast route, the CV bit is part of a new MVPN Flag EC, to be specified in a future revision.

2.5.4. Setting Up Forwarding State on RBRs

As a RBR follows the PMSI/Leaf A-D route procedures (even though the S-PMSI A-D route may be made up and not real), it sets up forwarding state accordingly [RFC7988] [RFC8556]. If IR is used in the upstream region, a downstream allocated label is advertised in the PTA of the Leaf A-D route sent upstream. If BIER is used in a region, the root RBR for the segment in that region MUST advertise an S-PMSI A-D route, whether the route is actually received from upstream or made up based on received C-multicast route or Leaf A-D route, with the PTA's label field set to a label upstream-assigned by the root RBR of the segment. This allows label switching by the RBR instead of relying on (C-S,C-G) lookup based forwarding in the VRF.

2.5.5. Other Types of Tunnels

The inter-region segmented tunnel can consists of different types of tunnels, like PIM/mLDP/RSVP-TE P2MP tunnels that require advertised S-PMSI A-D routes. This is just like BIER case mentioned in the above section. The only difference is that in BIER case it is the upstream allocated label that needs to be advertised by the S-PMSI A-D routes and in PIM/mLDP/RSVP-TE P2MP case it is the tunnel identity and optionally the upstream allocated label that need to be advertised by the S-PMSI A-D routes.

3. Security Considerations

This document does not seem to introduce new security risks, though this may be revised after further review and scrutiny.

4. Acknowledgements

The authors thank Vinay Nallamothu and Kevin Wang for their comments and suggestions. The authors also thank Vinod N Kumar and Sambasiva Rao for their suggestion of using the selected UMH route's RD for C-multicast A-D even when the source is not in the same AS (Section 1.4, Section 2.2.2).

5. References

5.1. Normative References

[I-D.ietf-bess-evpn-bum-procedure-updates]
Zhang, Z., Lin, W., Rabadan, J., Patel, K., and A. Sajassi, "Updates on EVPN BUM Procedures", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-bum-procedure-updates-14, 18 November 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-bum-procedure-updates-14.txt>.

[I-D.ietf-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", Work in Progress, Internet-Draft, draft-ietf-bess-mvpn-evpn-aggregation-label-08, 20 January 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-mvpn-evpn-aggregation-label-08.txt>.

[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>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <https://www.rfc-editor.org/info/rfc5015>.

[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>.

[RFC7524] Rekhter, Y., Rosen, E., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area Point-to-Multipoint (P2MP) Segmented Label Switched Paths (LSPs)", RFC 7524, DOI 10.17487/RFC7524, May 2015, <https://www.rfc-editor.org/info/rfc7524>.

[RFC7716] Zhang, J., Giuliano, L., Rosen, E., Ed., Subramanian, K., and D. Pacella, "Global Table Multicast with BGP Multicast VPN (BGP-MVPN) Procedures", RFC 7716, DOI 10.17487/RFC7716, December 2015, <https://www.rfc-editor.org/info/rfc7716>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[RFC7988] Rosen, E., Ed., Subramanian, K., and Z. Zhang, "Ingress Replication Tunnels in Multicast VPN", RFC 7988, DOI 10.17487/RFC7988, October 2016, <https://www.rfc-editor.org/info/rfc7988>.

[RFC8556] Rosen, E., Ed., Sivakumar, M., Przygienda, T., Aldrin, S., and A. Dolganow, "Multicast VPN Using Bit Index Explicit Replication (BIER)", RFC 8556, DOI 10.17487/RFC8556, April 2019, <https://www.rfc-editor.org/info/rfc8556>.

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

5.2. Informative References

[I-D.ietf-bess-evpn-irb-mcast]
Lin, W., Zhang, Z., Drake, J., Rosen, E. C., Rabadan, J., and A. Sajassi, "EVPN Optimized Inter-Subnet Multicast (OISM) Forwarding", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-irb-mcast-07, 23 June 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-irb-mcast-07.txt>.

[I-D.zzhang-bess-mvpn-evpn-segmented-forwarding]
Zhang, Z. and J. Xie, "MVPN/EVPN Segmentated Forwarding Options", Work in Progress, Internet-Draft, draft-zzhang-bess-mvpn-evpn-segmented-forwarding-00, 20 December 2018, <https://www.ietf.org/archive/id/draft-zzhang-bess-mvpn-evpn-segmented-forwarding-00.txt>.

[RFC9026] Morin, T., Ed., Kebler, R., Ed., and G. Mirsky, Ed., "Multicast VPN Fast Upstream Failover", RFC 9026, DOI 10.17487/RFC9026, April 2021, <https://www.rfc-editor.org/info/rfc9026>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks
Email: zzhang@juniper.net

Robert Kebler
Juniper Networks
Email: rkebler@juniper.net

Wen Lin
Juniper Networks
Email: wlin@juniper.net

Eric Rosen
Email: erosen52@gmail.com