Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 24 May 2024

F. Duan
S. Chen
Huawei Technologies
Y. Liu
China Mobile
H. Wang
Ruijie Networks Co., Ltd.

21 November 2023

Multicast VPN Upstream Designated Forwarder Selection
draft-wang-bess-mvpn-upstream-df-selection-08

Abstract

This document defines Multicast Virtual Private Network (VPN) extensions and procedures of designated forwarder election performed between ingress PEs, which is different from the one described in [RFC9026] in which the upstream designated forwarder determined by using the "Standby PE Community" carried in the C-Multicast routes. Based on the DF election, the failure detection point discovery mechanism between DF and standby DF is extended in MVPN procedures to achieve fast failover by using BFD session to track the status of detection point. To realize a stable "warm root standby", this document obsolete the P-Tunnel status determining procedure for downstream PEs in regular MVPN by introducing a RPF Set Checking mechanism as an instead.

Requirements Language

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

Status of This Memo

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

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

Copyright Notice

Table of Contents

1. Introduction

MVPN [RFC6513] and [RFC6514] defines the MVPN architecture and MVPN protocol specification which include the basic procedures for selecting the Upstream Multicast Hop. Further [RFC9026] defines some extensions to select the primary and standby upstream PE for a VPN multicast flow on downstream PEs. After selecting the Upstream Multicast Hop, the downstream PEs send MVPN C-Multicast routes to both primary and standby Upstream PE. Upon receiving the MVPN join routes, the upstream / ingress PEs can either perform "hot root standby" or "warm root standby". For the "hot root standby" mechanism, all the ingress PEs, regardless of the primary or standby role, forward (C-S,C-G) flow to other PEs through a P-tunnel, forcing the egress PEs to discard all but one. In this way, the failover can be conducted by leaf PE within extremely short duration when the failure of upstream link or device is detected. However, this will cause the steady traffic redundancy throughout the backbone network. In the scenario where bandwidth waste issue is concerned, such as enterprise networks crossing provider networks, the "warm root standby" mechanism is expected to be a better solution. However, there are some problems when deploying the "warm root standby" mechanism described in [RFC9026].

a. Upon the failure of primary ingress PE, the leaf PE needs to send the new C-multicast route towards the standby ingress PE without carrying the Standby PE BGP Community according to [RFC9026]. Leaf PE needs to update all relevant C-multicast routes and sends them to the standby ingress PE. For example, if there are 1000 (C-S,C-G)s, 1000 C-multicast routes will be updated and resent so that the standby PE can finally forward traffic. The failover time can hardly reach the same level of "hot root standby" mechanism.

b. There is no endogenous mechanism for standby ingress PEs to discover and detect the failure of primary ingress PEs, resulting in the uncertainty in deployment and implementation. If the standby ingress PE can directly detect the failure of the primary ingress PE, it can take over the role of designated forwarder and send the traffic immediately.

c. In [RFC9026], the standby ingress PE is determined by using "Standby PE Community" carried in the C-Multicast routes. The premise of this mechanism is that all leaf PEs choose the same primary and standby ingress PEs, which may not be met due to transient unicast routing inconsistencies, the inconsistencies of

P-Tunnel status determined by each leaf PE or lack of support of the Standby PE community on leaf PE, causing that the "warm root standby" mechanism is not stable and returns to "hot root standby" mode because the standby ingress PE also sends multicast traffic to backbone when the condition is not satisfied.

d. When the primary and standby designated forwarders are selected based on IP addresses of root PEs, the primary and standby roles are fixed for each multicast flow. Ingress PEs cannot perform load balancing for different multicast traffic. Hashing algorithm in [RFC6513] utilized source and group addresses and allows load balancing for different (C-S,C-G)s. However, the specific procedure of selecting a standby PE was not specified.

The hot root standby is good at fast failover. The warm root standby has advantages of saving the bandwidth. In order to have both advantages of hot root standby and warm root standby, this document defines a new MVPN procedure of designated forwarder election performed between ingress PEs. Based on the DF election, the failure detection point discovery mechanism between DF and standby DF is extended to achieve fast failover by using a BFD session to track the status of detection point. To realize a stable "warm root standby", this document obsoletes the P-Tunnel status determining procedure for downstream PEs in regular MVPN by introducing a RPF Set Checking mechanism as an instead.

2. Terminology

The terminology used in this document is the terminology defined in[RFC6513], [RFC6514] and [RFC9026].

For convenience of description, the abbreviations used in this document is listed below.

DF: Designated Forwarder

IDF: Ingress Designated Forwarder

UMH: Upstream Multicast Hop

P-tunnel: Provider-Tunnel

VPN: Virtual Private Network

MVPN: Multicast VPN

GTM: Global Table Multicast

RD: Route Distinguisher

NLRI: Network Layer Reachability Information

BFD: Bidirectional Forwarding Detection

MD: My Discriminator

VRI: VRF Route Import Extended Community

RR: Route Reflector

SFS: Selective Forwarder Selection

PTA: PMSI Tunnel Attribute

A new term is introduced below.

RPF Set Checking: RPF Set is a set of valid upstream interfaces that can accept multicast traffic. RPF Set checking allows multicast traffic to be received from backup P-Tunnel quickly when failure occurs.

3. Scenario

3.1. Passive IDF Negotiation Mode

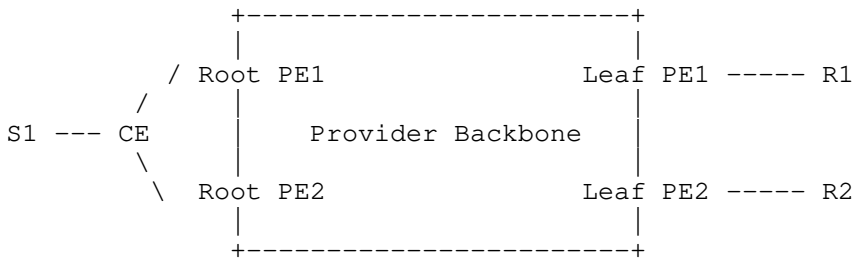

Figure 1: Passive IDF Negotiation Mode

In this scenario, the interfaces multihoming CE to provider's root PEs are bundled together and working in a eth-trunk mode, and a multichassis protocol is running between the multi-homed root PEs to coordinate with the CE to perform single active or all active data sending mode between CE and root PEs. Regardless either of the two sending mode is chosen, CE received multicast data from S1 only selects one interface to forward traffic, thus the root PE homed by

the selected interface is responsible for sending the corresponding multicast traffic to leaf PEs. The multi-homed root PEs do not really run an IDF negotiation procedure between themselves but accept the IDF role passively. Therefore, we call this scenario using Passive IDF Negotiation Mode in this document.

3.2. Active IDF Negotiation Mode

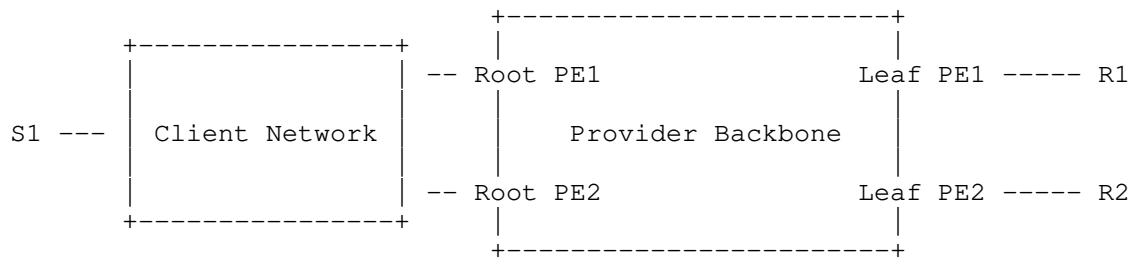

Figure 2: Active IDF Negotiation Mode

In this scenario, the "Client Network" is a layer 3 network area containing one or more CE routers. If only one CE router is included in the "Client Network" , the main difference between this circumstance and above is that the interfaces multihoming CE to root PEs are not bundled and each of them is an individual layer 3 interface. The IP subnet of the multihoming interfaces can be in either same or different, each of the multi-homed root PEs can receive one copy of the specific multicast stream (S1, G) received through the "Client Network". For the "warm root standby" mechanism, only one root PE (Called IDF in this document) can send the received multicast traffic to leaf PEs through provider's backbone. Thus the IDF must be selected among the multi-homed root PEs by themselves. So, in this document, we call this scenario using Active IDF Negotiation Mode.

4. Specification

4.1. IDF Negotiation Community

This community is carried in the UMH routes and used by the multi-homed root PEs to notify each other to perform IDF election. Leaf PEs can also check whether the UMH route is containing this community to perform checking according to the RPF Set Checklist. The value of this community will be allocated by IANA for each negotiation mode individually from the "Border Gateway Protocol (BGP) Well-known Communities" registry using the First Come First Served registration policy.

4.2. BFD Discriminator Attribute

This attribute is carried in UMH routes and its format reuses the one defined in [RFC9026] with the "BFD Mode" field redefined as a unicast BFD session type, of which the value is recommended to be 2 and will be allocated by IANA according to the registration policy. The source IP optional TLV in this document is mandatory and used to discover the failure detection point of the IDF.

5. Procedure

5.1. Signaling

In this section, the procedure is under the condition that the value of the RDs of multi-homed root PEs for a same MVPN are distinct, which means that the VPN route originated by each multi-homed PE can be received by the others and leaf PEs can also perform SFS reliably.

5.1.1. Originating VPN Routes to Multicast Sources

To perform IDF election procedure in this document, the multi-homed root PEs MUST include an IDF negotiation Community in the originating VPN routes to multicast sources. The negotiation mode (Passive or Active) is determined by the connection type of the Client network / CE, and MUST be configured consistently on each multi-homed root PE.

In order to perform endogenous mechanism of IDF election and fast failure detection, the BFD Discriminator Attribute described in section 4.2 MUST also be carried when each multi-homed root PE originates a UMH routes, with the MD field filled with a local configured BFD discriminator and the IP address field of the Source IP TLV filled with the local IP of the interface connecting to the Client network / CE, from which the prefix of the originating UMH route is learned. If the UMH prefix is learned from more than one local interface, the one chosen to fill the Source IP TLV of the BFD Discriminator Attribute MUST be consistent with the one selected as RPF interface for the multicast stream sent by the corresponding multicast source. In this document, the filled Source IP address is the failure detection point, if the corresponding root PE is selected as the IDF of a specific multicast stream, it is used to establish a BFD session to do fast tracking of failure of IDF. In IPv6 scenarios, a global IPv6 address SHOULD be configured on the client facing interfaces to succeed in the establishment of multi-hop IPv6 BFD sessions.

5.1.2. Originating C-Multicast Routes

If a leaf PE decides to send C-Multicast routes to upstream PEs for a given (C-S,C-G), it follows the procedure described in [RFC6514] excepting that the RPF route of the c-root has an IDF negotiation community. According to the negotiation community, a distinct C-Multicast route for (C-S,C-G) is sent to each multi-homed root PE. Leaf PE installs all P-Tunnels rooted from the multi-homed PEs into the RPF tunnel checklist of the corresponding multicast traffic (C-S,C-G).

If there is a local receiver connected to one of the multi-homed root PEs and the Passive IDF Negotiation Mode is performed between them, the root PE having local receivers sends the specific C-Multicast route (C-S,C-G) joined by the local receivers to the multi-homed others, after which it installs all P-Tunnels rooted from the multi-homed others and local upstream interface into the RPF tunnel checklist of the corresponding multicast traffic (C-S,C-G).

5.1.3. Ingress Designated Forwarder Selection

For Passive IDF election, it is performed by CE routers as described in section 3.1. This section describes two optional solutions for Active IDF election.

5.1.3.1. Out-Of-Band Mechanism

VRRP specifies an election protocol that dynamically assigns responsibility of a virtual router to one of the VRRP routers on a LAN. The VRRP router controlling the IPv4 or IPv6 address(es) associated with a virtual router is called the Master, and it forwards packets sent to these IPv4 or IPv6 addresses. Similarly, the role of the VRRP routers associated with a virtual router can also be that of the upstream PEs in MVPN dual homing upstream PEs deployment.

The method of mapping the role of a VRRP router to that of a MVPN upstream PE is more likely an administrative measure and could be implemented as configurable policies. Both the primary and standby PEs install VRF PIM state corresponding to BGP Source Tree Join route and send C-Join messages to the CE toward C-S. Whereas only the primary upstream PE (Virtual Router Master according to VRRP) forwards (C-S,C-G) flow to downstream PEs through a P-tunnel if IDF election is performing between the upstream PEs.

Other private implementations or similar designated forwarder selection technologies could also be optional. However, a feasible technology should has the ability to be deployed per VRF and be associated with one Multicast VPN instance. All PEs connected to the same customer's layer 3 network area MUST keep a coincident status of whether performing IDF election or not by negotiating dynamically or being configured manually, the dynamic protocol for negotiation of this status is outside the scope of this document.

5.1.3.2. Endogenous Mechanism

Considering a multicast source connecting to the client network area multihoming to the provider network, the prefix of the source can be learned by all multi-homed root PEs, each of which originates a corresponding VPN route with a VRI Extended Community including the originator's IP address to the others and leaf PEs. According to that, each multi-homed root PE can learn all the others' originator IP addresses for a specific multicast source, based on which the IDF can be calculated consistently on each root.

The default procedure for IDF election is at the granularity of (C-S,C-G). There are two options listed below for IDF election of a specific multicast source C-S, a deployment can use each of them and MUST be configured consistently among the multi-homed root PEs:

a. To perform single IDF election for all C-Gs of a specific multicast source C-S, each PE builds an ordered list in ascending order of the IP addresses of all multi-homed PE nodes learning the UMH routes to the multicast source C-S (including itself). As described in the first paragraph of this section, each IP address in this list is extracted from the Global Administrator field of VRI Extended Community carried in those UMH routes related to the specified multicast source C-S. Every PE is then given an ordinal indicating its position in the ordered list, starting with 0 as the ordinal for the PE with the numerically lowest IP address. The originator IP address with ordinal 0 is the winner, and the corresponding root PE is selected as IDF by every PE. The root PE of which the corresponding originator IP address is sub-optimal is selected as Standby IDF.

b. To perform IDF election for each C-G of a specific multicast source C-S, each PE also builds an ordered list of the IP addresses of all the multi-homed PE nodes at first. The difference between this option and above is that the election of IDF occurs not upon receiving all UMH routes of the other multi-homed PEs of the specified C-S but upon receiving the C-multicast join of the corresponding C-G. Assuming an ordered list of N elements, the PE with ordinal i is the IDF for a C-G when (C-G

mod N) = i. The PE with ordinal j is the Standby IDF when j is (C-G mod (N-1)). The calculation of standby IDF uses the ordered IP addresses list without considering the existence of the elected IDF element.

In order to reduce traffic waste between the Client Network and root PEs, a root PE can only send C-PIM Join messages towards the Client Network if it is the primary or standby DF.

5.1.4. Failure Detection and Fast Failover

For the Passive IDF Negotiation Mode, the CE router is responsible for the failure detection of multihoming links or multi-homed PE nodes using some existing solution, which is out the scope of this document. For the Active IDF Negotiation Mode with Out-Of-Band Mechanism described in section 5.1.3.1, the failure detection solution is always built in the multichassis protocols used for IDF election. This section only details the failure detection and fast failover procedure for the Active IDF negotiation mode with endogenous mechanism. Two methods are proposed to detect the failure.

5.1.4.1. BFD Method

To detect the failure of the node or the client facing link of IDF quickly, after the election of IDF PE and Standby IDF PE, the Standby IDF initializes a BFD session. Several important parameters of the BFD session are introduced as follows. The source IP of the BFD session uses a local configured IP address of the corresponding multicast VRF. The destination IP is extracted from the Source IP TLV of BFD Discriminator Attribute carried in the UMH route sent by the IDF. MD is filled with the MD field of BFD Discriminator Attribute carried in VPN routes originated by current Standby IDF. The YD(Your Discriminator) of the BFD session is dynamically learned through the BFD initialization procedure.

Upon the occasion of the failure, the status of the BFD session goes down. The Standby IDF PE of the C-Gs selecting the failure / affected node as IDF takes over the primary role and sends the multicast traffic belonging to C-Gs to leaf PEs through the backbone. The failure / affected PE withdraws its VPN route advertised before, this will re-trigger the procedure described in section 5.1.3.2 and a new IDF PE (which was the old Standby IDF PE) and Standby IDF PE will be selected. The new standby IDF MUST send C-PIM Join message towards Client Network to receive multicast traffic.

If the previous failure node / link goes up again or a new multi-homed PE of the specified multicast source is coming up and the IDF PE is calculated to be changed, the new IDF will take over the running IDF. To avoid data transfer crash, the running IDF (That should be the new Standby IDF) does not trigger the establishment of BFD session with new IDF until the local configured failback time expires, during which it keeps the IDF role and waits the new IDF completing the establishment of the multicast path from the SDR of the specified multicast source to itself. Upon the occasion of BFD session goes up, the running IDF stops sending multicast traffic to leaf PEs and the new IDF takes over the IDF role to send multicast stream for (C-S,C-G).

5.1.4.2. Monitoring traffic from IDF

The second method needs standby IDF to detect the failure by joining a P-Tunnel rooted at the IDF and monitoring the traffic received from the P-Tunnel. Even though there may be no local receiver connected to the standby IDF, the standby IDF needs to join the P-Tunnel by sending Leaf A-D Route or P-Tree Signaling. Standby IDF then sends the C-Multicast Route to IDF in order to receive traffic from the P-Tunnel. IDF will receive traffic from the IDF and the client facing interface simultaneously. However, it does not forward traffic to leaf PE when failure is not detected. When failures occur between the client network and IDF, the standby IDF will no longer receive any traffic from the P-Tunnel. The detection of interrupted flow will trigger the role transition from standby IDF to IDF. Then the new IDF will forward traffic to leaf PE.

However, the standby IDF may also cannot receive traffic when the failure occurs between the IDF and standby IDF. Under this circumstance, the standby IDF will switch to IDF when the client facing link and IDF still work well. There will be dual IDFs and leaf PE will receive two copies of the same flow. Suggestions about deployment are provided to avoid this situation:

a. Multiple parallel links are suggested to be deployed between the IDF and standby IDF. The probability of dual IDFs due to link failure can be greatly reduced.

b. PMSI tunnel protection can be utilized together. When the link between IDF and standby IDF fails, the underlay local protection of PMSI Tunnel can ensure that standby IDF can still receive traffic from IDF and avoid the dual-IDF situation.

5.2. Data Forwarding

5.2.1. Procedure on Root PEs

For the Passive IDF Negotiation Mode, the set of leaves of P-Tunnel rooted at each multi-homed PE has the others as members if the others have local receivers willing to accept the corresponding C-Flow. The detailed signaling procedure is described in section 5.1.2. When CE sends multicast data performing load balance to only one root PE (Which is the Passive IDF), IDF send this multicast traffic to the leaf PEs and the other multi-homed root PEs. When the multi-homed root PEs receive the C-Flow, it MUST perform RPF Set Checking, by accepting the data from either the client facing interface learning the corresponding route of the multicast source or anyone of the P-Tunnels rooted at the other multi-homed PEs. To avoid multicast traffic loop and duplication, the data received from the P-Tunnels at each root PE MUST NOT send back to P-Tunnels again and can only be forwarded to the local receivers of the receiving PE.

For the Active IDF Negotiation Mode, each multi-homed root PE receives a copy of C-Flow and forwards the multicast traffic to its local receivers. Only DF can send data to leaf PEs through backbone. All of the multi-homed root PEs perform RPF Set checking by matching their client facing interface exactly.

5.2.2. Procedure on Leaf PEs

For either of the two IDF negotiation modes described in this document, leaf PEs install each P-Tunnel rooted at each multi-homed root PE into the RPF Set checklist for the corresponding multicast flow (C-S,C-G), thus the multicast data sent by each of the multi-homed root PEs can be accepted by leaf PEs. Upon the failure of IDF, the Standby IDF takes over the primary role and leaf PEs are ready to receive the data sent by the new primary IDF with no latency thanks to the RPF Set checking mechanism.

5.3. Distinguishing UMH and C-multicast Routes

It was recommended in RFC 6514, on each multi-homed root PE, the UMH VRF of the MVPN MUST use its own distinct RD to support non-congruent unicast and multicast connectivity, the procedure described in above section is also under this premise. However, in [RFC7716], the UMH routes are not sent in the VPN-IP SAFI and there is no RD included in the NLRI key. There are also some other scenarios that the UMH VRF of the MVPN on the multi-homed PEs MUST be configured with a same RD for some deployment reasons, which causing that the IDF negotiation procedure can hardly be performed because that the UMH route originated by each multi-homed root PE cannot be collected reliably

by the other root PEs and leaf PEs because of the route selecting mechanism on BGP RRs. When UMH routes to same multicast source from different root PEs carry same RD or no RD, they will be same from the perspective of leaf PEs. Because Originating Router's IP Address is not the key field when the UMH route is processed.

For the scenarios of the same RD, this document introduces a new type of UMH route to be sent in MVPN SAFI, of which the NLRI key consists of the following fields:

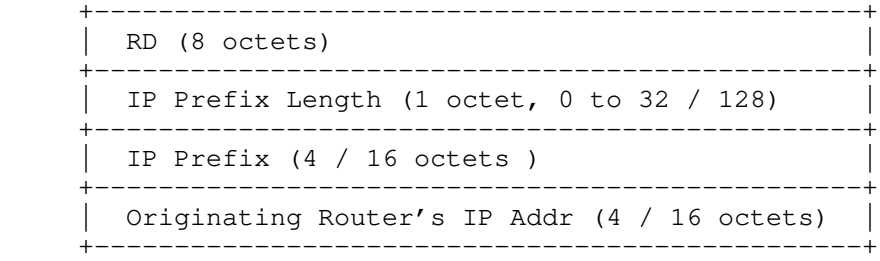

Figure 3: MVPN UMH Routes

The length of the IP Prefix field is determined by the address family of MVPN. If IPv4 is being used, it will be 4 octets. Otherwise it will be 16 octets for IPv6. After determining the length of IP Prefix field, the length of the Originating Router's IP Addr field is judged by NLRI key length. The type of this route will be allocated in IANA.

If the RDs of the UMH VRFs on the multi-homed root PEs are same, the root PEs import the routes of the client multicast sources to their local UMH VRFs and send above UMH routes to all other PEs of the MVPN. The UMH routes will carry a VRI Extended Community described in [RFC6514], an IDF negotiation Community and a BFD Discriminator Attribute described in this document. All the procedure applied to the VPN-IP routes described in [RFC6513] and [RFC6514] SHOULD be inherited by this UMH route. The receivers (which should be MVPN PEs) of this route MUST install it into their local multicast RIB as UMH route and it has a higher priority than other existing UMH route type while a MVPN PE using it to determine the upstream PE of a specified (C-S,C-G) or (C-*,C-G). The Originating Router's IP Addr will be used to identify UMH routes from different upstream PEs.

For the non-segmented Inter-AS P-Tunnel over IPv6 infrastructure scenarios, the length of Source AS field of C-Multicast routes cannot hold an IPv6 address, causing that it is hard to distinguish the two C-Multicast routes with a same granularity of (C-S,C-G) or (C-*,C-G)

sent to two ingress PEs individually. To solve this problem, this document introduces a Root Distinguisher Extended Community, which is an IP-address-specific Extended Community with a fixed type of IPv4. The Global Administrator field of this Extended Community is filled with a 4-octet global unique value configured. This 4-octet value and the IPv6 Originating Router's IP Addresses of each MVPN PE needs not to be a routable IPv4 address. The Local Administrator field of the Extended Community is filled with 0. The type and sub type of this Extended Community will be allocated in IANA.

The Root Distinguishing Extended Community is carried in the Intra-AS AD routes or the wildcard S-PMSI AD routes. According to [RFC6514] and [RFC6515], the non-segmented Inter-AS and IPv6 infrastructure scenarios are determined on MVPN leaf PEs. The Source AS field of the C-Multicast routes will be filled with the root distinguishing value of root PEs which the route is sent to.

5.4. Segmented Inter-AS Scenario

In the regular procedure of [RFC6514], Intra-AS AD route is only used in non-segmented Inter-AS scenario. In the segmented Inter-AS scenario, different Intra-AS AD routes originated by different PEs in the same AS are aggregated to a single Inter-AS AD route on ASBRs with the granularity of <AS, MVPN>. The specific original root PE's information is substituted with source AS during the aggregation, which results in that leaf PEs located in downstream ASes cannot differentiate two multicast traffic sent by different root PEs in the same original AS.

In this document, two approaches are proposed to facilitate the root PE selection of leaf PEs in downstream ASes.

This first approach is to use the wildcard S-PMSI AD route described in [RFC6625] instead of Intra-AS AD route. As described in [RFC6514], the S-PMSI AD route will not be aggregated by ASBR while being used to set up Inter-AS segmented S-PMSI tunnels, result in that Leaf PE in downstream AS can do explicit tracking of those tunnels established from the redundant PEs located in upstream AS. The propagation procedure between ASes follows the description in section 12.2 of [RFC6514].

The second method is to use PE Distinguisher Labels attribute defined in [RFC6514] to carry PE address and corresponding upstream assigned label at the segmentation point. In the segmented scenario, the PE Distinguisher Labels attribute SHOULD be distributed with the Inter-AS A-D route. When the multicast traffic is received from the Intra-AS P-Tunnel from an ingress root PE on ASBR, the ASBR will switch Intra-AS P-Tunnel to Inter-AS P-Tunnel and add the corresponding PE

Distinguisher Label as an inner label into the label stack. In this way, leaf PE will recognise the root PE of the multicast traffic and RPF Checking can be performed accordingly. As described in [RFC6514], leaf PE will find corresponding Inter-AS A-D route while sending C-multicast route. It fills source-AS field of the C-multicast route with the corresponding PE Distinguisher Label carried in the Inter-AS A-D route.

6. Backward Compatibility

When some devices do not support MVPN Upstream IDF Election, the following procedures are introduced to support the backward compatibility and end-to-end MVPN service. This section mainly discuss the situation that root PE and leaf PE does not support the IDF election function simultaneously.

6.1. Root PE Not Support IDF Election

When a root PE does not support MVPN Upstream IDF Election, the UMH route sent by the root PE will not carry IDF Negotiation Community. Other root PEs will check all received UMH routes with the same prefix. As long as one of these routes does not carry the community, the IDF election procedures will not be executed and hot root standby will be conducted.

Leaf PE will also check the received UMH routes. When one of the received UMH route does not carry the IDF Negotiation Community, the RPF Set checklist will not be used. The RPF Checking will be based on the normal procedure that only one upstream interface will be considered as the valid upstream interface for certain (C-S,C-G).

6.2. Leaf PE Not Support IDF Election

When leaf PE does not support the IDF Election or it cannot become the leaf of PMSI Tunnel rooted at the main IDF, it must revert back to join the normal PMSI P-Tunnel to receive multicast traffic. A new extra PMSI attribute called "Secondary PMSI Tunnel Attribute" is appended after the existing PTA in the x-PMSI A-D route sent by the primary and standby IDF to identify the warm PMSI Tunnel. The format and content of "Secondary PMSI Tunnel Attribute" are same as PTA defined in [RFC6514]. The attribute type (Attr Type) field will be allocated by IANA.

When leaf PE does not support the IDF Election function, it cannot recognise these Secondary attribute. Therefore, the leaf PE will join the PMSI Tunnel identified by the normal PTA defined in [RFC6514]. The explicit tracking defined in [RFC6514] will be conducted. Even if leaf PE supports the IDF Election function, it

may not be able to join the "Warm Tunnel" due to certain reasons, such as local policy. Under this condition, leaf PE can also join the original PMSI Tunnel.

When leaf PE decide to join the "Warm PMSI Tunnel" identified by the Secondary PTA, it will send C-Multicast route to root PEs. The x-PMSI Leaf A-D route will carry a "Secondary Indication Community". The value of this community will be allocated by IANA.

When root PEs are performing the IDF Election, only IDF can forward corresponding traffic into the PMSI Tunnel identified by "Secondary PTA". One multicast traffic can be carried by both the normal PMSI Tunnel and warm PMSI Tunnel simultaneously. Only the warm PMSI Tunnel is controlled by the aforementioned IDF Negotiation Status.

7. Security Considerations

This document follows the security considerations specified in [RFC6513] and [RFC6514]. In addition, because the establishment of segmented Inter-AS PMSI tunnel is introduced by using Intra-AS AD routes in this document, the Originator's IP addresses are exposed between ASes which may cause some security risks in the scenarios of different service providers for different ASes. In order to reduce the impact, the Intra-AS AD routes to be leaked between ASes MUST be controlled under security policies so that the numbers of the leaked Originator's IP addresses can be reduced.

8. IANA Considerations

This document defines a new BGP Community called IDF negotiation Community, of which the value will be allocated from IANA for each negotiation mode individually. The BFD Discriminator Attribute defined in [RFC9026] is reused and the value of BFD Mode is recommended to be 2 in this document, which will be reviewed by IANA.

This document defines a new UMH route type for MVPN, of which the value is recommended to be 8 and will be reviewed by IANA. This document defines a new BGP Extended Community called "Root Distinguisher", this Community is of an extended type and is transitive, the Type and Sub-Type are TBD and will be allocated from IANA.

This document defines a new PMSI Tunnel Attribute called "Secondary PMSI Tunnel Attribute" and a new community called "Secondary Indication Community". Their attribute types will be allocated by IANA.

9. Acknowledgements

The authors wish to thank Jingrong Xie and Jeffrey Zhang, for their reviews, comments and suggestions.

10. 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>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC5798] Nadas, S., Ed., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC 5798, DOI 10.17487/RFC5798, March 2010, <https://www.rfc-editor.org/info/rfc5798>.

[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>.

[RFC6515] Aggarwal, R. and E. Rosen, "IPv4 and IPv6 Infrastructure Addresses in BGP Updates for Multicast VPN", RFC 6515, DOI 10.17487/RFC6515, February 2012, <https://www.rfc-editor.org/info/rfc6515>.

[RFC6625] Rosen, E., Ed., Rekhter, Y., Ed., Hendrickx, W., and R. Qiu, "Wildcards in Multicast VPN Auto-Discovery Routes", RFC 6625, DOI 10.17487/RFC6625, May 2012, <https://www.rfc-editor.org/info/rfc6625>.

[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>.

[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>.

[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

Fanghong Duan
Huawei Technologies
Email: duanfanghong@huawei.com

Siyu Chen
Huawei Technologies
Email: chensiyu27@huawei.com

Yisong Liu
China Mobile
Email: liuyisong@chinamobile.com

Heng Wang
Ruijie Networks Co., Ltd.
Email: wangheng1@ruijie.com.cn