Bidirectional Forwarding Detection (BFD) for Multi-point Networks and Protocol Independent Multicast - Sparse Mode (PIM-SM) Use Case
draft-ietf-pim-bfd-p2mp-use-case-02

Abstract

This document discusses the use of Bidirectional Forwarding Detection (BFD) for multi-point networks to provide nodes that participate in Protocol Independent Multicast - Sparse Mode (PIM-SM) with the sub-second convergence. Optional extension to PIM-SM Hello, as specified in RFC 7761, to bootstrap point-to-multipoint BFD session. Also defined in this document.

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 January 26, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must
1.  Introduction

Faster convergence in the control plane, in general, is beneficial and allows minimizing periods of traffic blackholing, transient routing loops, and other scenarios that may negatively affect service data flow. That equally applies to unicast and multicast routing protocols.

[RFC7761] is the current specification of the Protocol Independent Multicast - Sparse Mode (PIM-SM) for IPv4 and IPv6 networks. Confirming implementation of PIM-SM elects a Designated Router (DR) on each PIM-SM interface. When a group of PIM-SM nodes is connected to shared-media segment, e.g., Ethernet, the one elected as DR is to act on behalf of directly connected hosts in the context of the PIM-SM protocol. Failure of the DR impacts the quality of the multicast services it provides to directly connected hosts because the default failure detection interval for PIM-SM routers is 105 seconds. Introduction of Backup DR (BDR), proposed in [I-D.ietf-pim-dr-improvement], improves convergence time in the PIM-SM over shared-media segment but still depends on long failure detection interval.

Bidirectional Forwarding Detection (BFD) [RFC5880] had been originally defined to detect failure of point-to-point (p2p) paths - single-hop [RFC5881], multihop [RFC5883]. [RFC8562] extends the BFD base specification [RFC5880] for multipoint and multicast networks, which precisely characterizes deployment scenarios for PIM-SM over LAN segment. This document demonstrates how point-to-multipoint (p2mp) BFD can enable faster detection of PIM-SM router failure and...
thus minimize multicast service disruption. The document also defines the extension to PIM-SM [RFC7761] and [I-D.ietf-pim-dr-improvement] to bootstrap a PIM-SM router to join in p2mp BFD session over shared-media link.

1.1. Conventions used in this document

1.1.1. Terminology

BFD: Bidirectional Forwarding Detection
BDR: Backup Designated Router
DR: Designated Router
p2mp: Point-to-Multipoint
PIM-SM: Protocol Independent Multicast - Sparse Mode

1.1.2. Requirements Language

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

2. Problem Statement

[RFC7761] does not provide a method for fast, e.g., sub-second, failure detection of a neighbor PIM-SM router. BFD already has many implementations based on HW that are capable of supporting multiple sub-second sessions concurrently.

3. Applicability of p2mp BFD

[RFC8562] may provide an efficient and scalable solution for the fast-converging environment that demonstrates the head-tails relationship. Each such group presents itself as p2mp BFD session with its head being the root and other routers being tails of the p2mp BFD session. Figure 1 displays the new optional BFD Discriminator TLV to bootstrap tail of the p2mp BFD session.
Figure 1: BFD Discriminator TLV to Bootstrap P2MP BFD session

where new fields are interpreted as:

<table>
<thead>
<tr>
<th>OptionType</th>
<th>OptionLength</th>
</tr>
</thead>
<tbody>
<tr>
<td>My Discriminator</td>
<td></td>
</tr>
</tbody>
</table>

OptionType is a value (TBA1) assigned by IANA Section 4 that identifies the TLV as BFD Discriminator TLV;

OptionLength value is always 4

My Discriminator - My Discriminator value allocated by the root of the p2mp BFD session.

If PIM-SM routers that support this specification are configured to use p2mp BFD for faster convergence, then the router to be monitored, referred to as ‘head’, MUST create BFD session of type MultipointHead, as defined in [RFC8562]. If the head doesn’t support [I-D.ietf-pim-dr-improvement], but, for example, uses procedures defined in [I-D.mankamana-pim-bdr], then it MUST include BFD TLV in its PIM-Hello message. If the head uses extensions defined in [I-D.ietf-pim-dr-improvement], then DR MUST include BFD TLV in its Hello message. The DR Address TLV also MUST be included in the Hello message. For a BDR it is RECOMMENDED to include BFD TLV in its Hello message. If BDR includes BFD TLV, then the BDR Address TLV also MUST be present in the Hello message. Then the head MUST begin periodic transmission of BFD control packets. Source IP address of the BFD control packet MUST be the same as the source IP address of the PIM-Hello with BFD TLV messages being transmitted by the head. The values of My Discriminator in the BFD control packet and My Discriminator field of the BFD TLV in PIM-Hello, transmitted by the head MUST be the same. When a PIM-SM router is configured to monitor the head by using p2p BFD, referred to through this document as ‘tail’, receives PIM-Hello packet with BFD TLV it MAY create p2mp BFD session of type MultipointTail, as defined in [RFC8562].

Because p2mp BFD doesn’t use the three-way handshake and the head transmits BFD control packets with the value of Your Discriminator field set to zero, [RFC8562] modified how a BFD system demultiplexes received BFD control packet. The tail demultiplexes p2mp BFD test session based on head’s source IP address, the My Discriminator value it learned from BFD Discriminator TLV and the identity of the
multipoint path that the BFD control packet was received from. The Detection Time for p2mp BFD sessions is defined differently from the definition provided in [RFC5880]. The Detection Time for each MultipointTail session is calculated as the product of the last received values of Desired Min TX Interval and Detect Mult. A tail declares the BFD session down after the Detection Timer expires. If the tail has detected MultipointHead failure, it MUST remove the neighbor. If the failed head node was PIM-SM DR or BDR, the tail MAY start DR Election process as specified in Section 4.3.2 [RFC7761] or Section 4.1 [I-D.ietf-pim-dr-improvement] respectively.

If the head ceased to include BFD TLV in its PIM-Hello message, tails MUST close the corresponding MultipointTail BFD session. Thus the tail stops using BFD to monitor the head and reverts to the procedures defined in [RFC7761] and [I-D.ietf-pim-dr-improvement].

3.1. Multipoint BFD Encapsulation

The MultipointHead of p2mp BFD session when transmitting BFD control packet:

MUST set TTL value to 1;

SHOULD use group address ALL-PIM-ROUTERS (‘224.0.0.13’ for IPv4 and ‘ff02::d’ for IPv6) as destination IP address

MAY use network broadcast address for IPv4 or link-local all nodes multicast group for IPv6 as the destination IP address;

MUST set destination UDP port value to 3784 when transmitting BFD control packets, as defined in [RFC8562].

4. IANA Considerations

IANA is requested to allocate a new OptionType value from PIM Hello Options registry according to:

<table>
<thead>
<tr>
<th>Value Name</th>
<th>Length Number</th>
<th>Name Protocol</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBA</td>
<td>4</td>
<td>BFD Discriminator</td>
<td>This document</td>
</tr>
</tbody>
</table>

Table 1: BFD Discriminator option type
5. Security Considerations

Security considerations discussed in [RFC7761], [RFC5880], and [RFC8562], and [I-D.ietf-pim-dr-improvement] apply to this document.

An implementation that supports this specification SHOULD use a mechanism to control the maximum number of BFD sessions that can be active at the same time.

6. Acknowledgments

Authors cannot say enough to express their appreciation of comments and suggestions we received from Stig Venaas.

7. Normative References

[I-D.ietf-pim-dr-improvement]

[I-D.mankamana-pim-bdr]


Authors' Addresses

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Ji Xiaoli
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
China

Email: ji.xiaoli@zte.com.cn
PIM Null register packing
draft-ietf-pim-null-register-packing-03

Abstract

In PIM-SM networks PIM registers are sent from the first hop router to the RP (Rendezvous Point) to signal the presence of Multicast source in the network. There are periodic PIM Null registers sent from first hop router to the RP to keep the state alive at the RP as long as the source is active. The PIM Null register packet carries information about a single Multicast source and group. This document defines a standard to send multiple Multicast source and group information in a single pim Null register packet and the interoperability between the PIM routers which do not understand the packet format with multiple Multicast source and group details.

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 October 18, 2019.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents
1. Introduction

PIM Null registers are sent by First hop routers periodically for Multicast streams to keep the states active on the RP as long as the Multicast source is alive. As the number of multicast sources increases, the number of PIM Null register packets that are sent increases at a given time. This results in more PIM packet processing at RP and FHR. The control plane policing (COPP), monitors the packets that gets processed by the control plane. Due to the high rate at which Null registers are received at the RP, this can lead to COPP drops of Multicast PIM Null register packets. This draft proposes a method to efficiently pack multiple PIM Null registers and register stop into a single message as these packets anyway don’t contain data. The draft also proposes interoperability with the routers that do not understand the new packet format.

1.1. Conventions Used in This Document

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].
1.2. Terminology

RP: Rendezvous Point
RPF: Reverse Path Forwarding
SPT: Shortest Path Tree
FHR: First Hop Router, directly connected to the source
LHR: Last Hop Router, directly connected to the receiver

2. PIM Register Stop format with capability option

A router (FHR) can decide to pack multiple Null registers based on the capability received from the RP as part of Register Stop. This ensures compatibility with routers that don’t support processing of the new format. The capability information can be indicated by the RP via the PIM register stop message sent to the FHR. Thus a FHR will switch to the new format only when it learns RP is capable of handling the packed Null register messages. Conversely, a FHR that doesn’t support the new format can continue generating the PIM Null register the current way. To exchange the capability information in the Register Stop message, the "reserved" field can be used to indicate this capability in those register stop messages. One bit of the reserved field is used to indicate the "packing" capability (P bit). The rest of the bits in the "Reserved" field will be retained for future use.

Figure 1: PIM Register Stop message with capability option

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5 6 7 8 9</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

PIM Version, Reserved, Type, Checksum, Group Address, Source Address
Same as RFC 7761 (Section 4.9.4)

P  Capability bit used to indicate support for Packed Null Register
3. New PIM Null register message

New PIM Null register message format includes a count to indicate the number of Null register records in the message.

Figure 2: New PIM Null Register message format

<table>
<thead>
<tr>
<th>PIM Ver</th>
<th>Type</th>
<th>SubType</th>
<th>Rsvd</th>
<th>Checksum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>+-------------------------------</td>
</tr>
<tr>
<td>count</td>
<td></td>
<td></td>
<td></td>
<td>+-------------------------------</td>
</tr>
<tr>
<td>Group Address[1] (Encoded-Group format)</td>
<td></td>
<td></td>
<td>+-------------------------------</td>
<td></td>
</tr>
<tr>
<td>Source Address[1] (Encoded-Unicast format)</td>
<td></td>
<td></td>
<td>+-------------------------------</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td>+-------------------------------</td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td>+-------------------------------</td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td>+-------------------------------</td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td>+-------------------------------</td>
</tr>
<tr>
<td>Group Address[N]</td>
<td></td>
<td></td>
<td></td>
<td>+-------------------------------</td>
</tr>
<tr>
<td>Source Address[N]</td>
<td></td>
<td></td>
<td></td>
<td>+-------------------------------</td>
</tr>
</tbody>
</table>

PIM Version, Reserved, Checksum
Same as RFC 7761 (Section 4.9.3)

Type, SubType
The new packed Null Register Type and SubType values TBD

count
The count of the number of packed Null register records.
A record consists of Group and Source Address

Group Address
IP address of the Multicast Group

Source Address
IP Address of the Multicast Source

4. New PIM Register Stop message format

The new PIM register stop is message includes a count to indicate the number of records that are present in the message.
### Figure 3: New PIM Register Stop message format

<table>
<thead>
<tr>
<th>0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1</th>
<th>PIM Ver</th>
<th>Type</th>
<th>SubType</th>
<th>Rsvd</th>
<th>Checksum</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reserved2</td>
</tr>
<tr>
<td>count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reserved2</td>
</tr>
<tr>
<td>Group Address[1] (Encoded-Group format)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Address[1] (Encoded-Unicast format)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group Address[N]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Address[N]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PIM Version, Reserved, Checksum
Same as RFC 7761 (Section 4.9.3)

**Type**

The new Register Stop Type and SubType values TBD

**Record count**

The count of the number of packed register stop records.
A record consists of Group and Source Address

**Group Address**

IP address of the Multicast Group

**Source Address**

IP Address of the Multicast Source

---

5. Protocol operation
The following combinations exist -

FHR and RP both support the new PIM Register formats -
   a. FHR sends the PIM register towards the RP when a new source is detected
   b. RP sends a modified register stop towards the FHR that includes capability information by setting the P bit (Figure 2)
   c. Based on the receipt of new Register Stop, FHR will start packing of Null registers using the new packed register format (Figure 1)
   d. RP processes the new Null register message and can generate new register Stop messages by packing multiple S,Gs towards the same FHR (Figure 3)

FHR supports but RP doesn’t support new PIM Register formats -
   a. FHR sends the PIM register towards the RP
   b. RP sends a normal register stop without any capability information
   c. FHR then sends Null registers in the old format

RP supports but FHR doesn’t support the new PIM Register formats -
   a. FHR sends the PIM register towards the RP
   b. RP sends a modified register stop towards the FHR that includes capability information
   c. Since FHR doesn’t support the new format, it sends Null registers in the old format

6. PIM Anycast RP considerations

   The new PIM register format should be enabled only if its supported by all PIM anycast RP members in the RP set for the RP address.

7. IANA Considerations

   This document requires the assignment of 2 new PIM message types for the packed pim register and pim register stop.

8. Acknowledgments

   The authors would like to thank Stig Venaas and Umesh Dudani for contributing to the original idea and also their very helpful comments on the draft.

9. References
9.1. Normative References


9.2. Informative References


Authors' Addresses

Vikas Ramesh Kamath
VMware
3401 Hillview Ave
Palo Alto CA 94304
USA

Email: vkamath@vmware.com

Ramakrishnan Chokkanathapuram Sundaram
Cisco Systems, Inc.
Tasman Drive
San Jose CA 95134
USA

Email: ramaksun@cisco.com

Raunak Banthia
Cisco Systems, Inc.
Tasman Drive
San Jose CA 95134
USA

Email: rbanthia@cisco.com
Multicast-Only Fast Reroute Based on Topology Independent Loop-free Alternate Fast Reroute
draft-liu-pim-mofrr-tilfa-00

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on December 26, 2019.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.
Abstract

Multicast-only Fast Reroute (MoFRR) has been defined in [RFC7431], but the selection of the secondary multicast next hop only according to the loop-free alternate fast reroute, which has restrictions in multicast deployments. This document describes a mechanism for Multicast-only Fast Reroute by using Topology Independent Loop-free Alternate fast reroute, which is independent of network topology and can achieve covering more network environments.

Table of Contents

1. Introduction ................................................ 2
   1.1. Requirements Language .................................. 3
   1.2. Terminology ............................................ 3
2. Problem Statement ........................................... 3
3. Solution .................................................... 4
   3.1. Secondary UMH Selection in PIM ........................... 5
   3.2. Secondary UMH Selection in MLDP .......................... 5
   3.3. Extension Protocol Fields Conflict ....................... 6
4. Packet Format ............................................... 6
   4.1. PIM Join Message Extension ............................... 7
   4.2. MLDP Label Mapping Message Extension ..................... 8
5. IANA Considerations ........................................ 11
6. Security Considerations ..................................... 11
7. References .................................................. 11
   7.1. Normative References .................................... 11
   7.2. Informative References .................................. 12
8. Acknowledgments ............................................. 12

1. Introduction

As the deployment of video services, operators are paying more and more attention to solutions that minimize the service disruption due to faults in the IP network carrying the packets for these services. Multicast-only Fast Reroute (MoFRR) has been defined in [RFC7431], which can minimize multicast packet loss in a network when node or link failures occur by making simple enhancements to multicast routing protocols such as Protocol Independent Multicast (PIM) and Multipoint LDP (mLDP). But the selection of the secondary multicast next hop only according to the loop-free alternate fast reroute in [RFC7431], and there are limitations in multicast deployments for this mechanism. This document describes a new mechanism for Multicast-only Fast Reroute using Topology Independent Loop-free Alternate (TILFA) fast reroute, which is independent of network topology and can achieve covering more network environments.
1.1. Requirements Language

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

1.2. Terminology

This document uses the terms defined in [RFC7431], and also uses the concepts defined in [RFC7490]. The specific content of each term is not described in this document.

2. Problem Statement

In [RFC7431] section 3, the secondary Upstream Multicast Hop (UMH) of PIM and mLDP for MoFRR is a loop-free alternate (LFA). However, the traditional LFA mechanism needs to satisfy at least one neighbor whose next hop to the destination node is an acyclic next hop, existing limitations in network deployments, and can only cover part of the network topology environments. In some network topology, the corresponding secondary UMH cannot be calculated, so PIM and mLDP cannot establish a standby multicast tree and cannot implement MoFRR protection. Therefore, the current MoFRR of PIM and mLDP is only available in the network topology applicable to LFA.

The remote loop-free alternate (RLFA) defined in [RFC7490] is extended from the LFA and can cover more network deployment scenarios through the tunnel as an alternate path. The RLFA mechanism needs to satisfy at least one node assumed to be N in the network that the fault node is neither on the path from the source node to the N node, nor on the path from the N node to the destination node. RLFA only has enhancement compared to LFA but still has limitations in network deployments.

[I-D.ietf-rtgwg-segment-routing-ti-lfa] defined a unicast FRR solution based on the TILFA mechanism. The TILFA mechanism can express the backup path with an explicit path, and has no constraint on the topology, providing a more reliable FRR mechanism. The unicast traffic can be forwarded according to the explicit path list as an alternate path to implement unicast traffic protection, and can achieve full coverage of various networking environments.

The alternate path provided by the TILFA mechanism is actually a Segment List, including one or more Adjacency SIDs of one or more links between the P space and the Q space, and the NodeSID of the P space node. PIM and mLDP can look up the corresponding node IP address in the unicast route according to the NodeSID, and the IP
addresses of the two endpoints of the corresponding link in the unicast route according to the Adjacency SIDs, but the multicast protocol packets cannot be directly sent along the path of the Segment List.

Both the PIM join message and the MLDP Label Mapping message need to be sent hop-by-hop to establish a standby multicast tree. However, not all of the nodes and links on the unicast alternate path are included in the Segment List. If the PIM and MLDP protocol packets are transmitted only in unicast mode, then equivalently the PIM and MLDP packets are transmitted through the unicast tunnel like unicast traffic, and cannot pass through the intermediate nodes of the tunnel. The intermediate nodes of the alternate path cannot forward multicast traffic because there are no PIM or MLDP state entries on the nodes. PIM needs to create entries on the device hop-by-hop and generate an incoming interface and an outgoing interface list. MLDP needs to create entries on the device hop-by-hop and generate an incoming label and an outgoing label list. So it can form an end-to-end complete multicast tree for forwarding multicast traffic. Therefore, it is not possible to send PIM and MLDP packets like unicast traffic according to the Segment List path and establish a standby multicast tree.

It is available in principle that the path information of the Segment List is added to the PIM and MLDP packets to guide the hop-by-hop RPF selection. The IP address of the node corresponding to the NodeSID can be used as the segmented root node, and the IP addresses of the interfaces at both endpoints of the link corresponding to the Adjacency SID can be used directly as the local upstream interface and upstream neighbor, but there is currently no field in protocol packet to carry the explicit path specified by the Segment List. For the PIM protocol, the PIM RPF Vector attribute was defined in [RFC5496], which can carry the node IP address corresponding to the NodeSID. The Explicit RPF Vector was defined in [RFC7891], which can carry the peer IP address corresponding to the Adjacency SID, but if there are multiple same peer IP addresses corresponding to the Adjacency SID (i.e. anycast IP address), the upstream neighbor of RPF selection may be different from the actual upstream link corresponding to the Adjacency SID, which can make the PIM join path and the TILFA calculation path inconsistent. For the MLDP protocol, there is also no field defined in the MLDP protocol Label Mapping message that can carry the explicit path of the Segment List.

3. Solution

An Upstream Multicast Hop (UMH) is a candidate next-hop that can be used to reach the root of the tree. In This document the secondary UMH is based on unicast routing to find Segment List calculated by
TILFA. With MoFRR, the procedures for determining the secondary UMH and establishing standby multicast tree are different for PIM and mLDP.

This document extends the PIM and mLDP protocol, to establish the standby multicast tree according to the Segment List calculated by TILFA, and can achieve full coverage of various networking environments for MoFRR protection of multicast services.

Assume that the Segment List calculated by TILFA is (NodeSID(A), AdjSID(A-B)). Node A belongs to the P Space, and node B belongs to the Q space. The IP address corresponding to NodeSID(A) can be looked up in the local link state database of the IGP protocol, and can be assumed to be IP-a. The IP addresses of the two endpoints of the link corresponding to AdjSID(A-B) can also be looked up in the local link state database of the IGP protocol, and can be assumed to be IP-La and IP-Lb.

3.1. Secondary UMH Selection in PIM

In the procedure of PIM, IP-a can be looked as the normal RPF vector attribute and added to the PIM join packet. IP-La and IP-Lb can be looked as the RPF Vector attribute of the adjacency relationship, called Adjacency RPF Vector, which is a new type of PIM join attributes, and added to the PIM join packet too.

The PIM protocol firstly can select the RPF incoming interface and upstream towards IP-a, and can join hop-by-hop to establish the PIM standby multicast tree until the node A. On the node A, IP-Lb can be looked as one PIM neighbor. If there are multiple PIM neighbors with the same address IP-Lb, all of the corresponding local interfaces on the node A need to be checked. The interface that is the only one with the IP address IP-La can be looked as the RPF incoming interface. The node A can send the PIM join packet to the node B on the interface of IP-La, and IP-Lb is used as the RPF upstream address of the PIM join.

After the PIM join packet is received on the node B, the PIM protocol can find no more join attributes and select the RPF incoming interface and upstream towards the multicast source directly, and then can continue to join hop-by-hop to establish the PIM standby multicast tree until the router directly connected the source.

3.2. Secondary UMH Selection in MLDP

In the procedure of MLDP, Explicit path TLV is newly defined in MLDP Label Mapping message to carry IP-a, IP-La and IP-Lb, which is contained in the field of Optional Parameters. IP-a can be looked as
the segmented root node address and is added as the Node Address Sub TLV in the Explicit path TLV. IP-La and IP-Lb are added as the Adjacency Address Sub TLV in the Explicit Path TLV.

The MLDP protocol can look up the upstream interface and the upstream LSR in the unicast route to IP-a, and can send the Label Mapping message hop-by-hop to establish the standby MPLS multicast tree to the node A. After the message is received on the node A, the Node Address Sub TLV corresponding to the IP-a can be deleted from the Label Mapping message.

On the node A IP-Lb can be looked as one MLDP neighbor. If there are multiple MLDP neighbors with the same address IP-Lb, all of the corresponding local interfaces on the node A need to be checked. The interface that is the only one with the IP address IP-La can be looked as the upstream interface. The node A can send MLDP Label mapping message to the node B, and IP-Lb is used as the upstream LSR address.

After the message is received on the node B, the Adjacency Address Sub TLV corresponding to the IP-La and IP-Lb is deleted from the Label Mapping message and if there is no more any sub TLV in the Explicit Path TLV then the TLV should be deleted. The MLDP protocol can select the upstream interface and the upstream LSR in the unicast route to the original root node directly, and can continue to send the Label Mapping message to establish the standby MPLS multicast tree to the original root node.

3.3. Extension Protocol Fields Conflict

PIM Adjacency RPF Vector attribute is newly defined in join attributes. If there are conflicts from multiple downstream PIM neighbors, the mechanism in [RFC5384] Section 3.3.3 can be used to select a PIM downstream neighbor with a numerically smallest IP address. If at least two neighbors have the same IP address, the interface index MUST be used as a tie breaker.

In the Explicit Path TLV newly defined in MLDP Label Mapping message, if there are conflicts from multiple downstream MLDP neighbors, including the inconsistency of the Sub TLV types, and the inconsistency of the Sub TLV contents, and the inconsistency of the Sub TLV sequences, it is also recommended to use the mechanism in [RFC5384] Section 3.3.3.

4. Packet Format

This section describes the format of PIM and mLDP protocol packet extension introduced by this document.
4.1. PIM Join Message Extension

The original PIM join attribute already has been defined in [RFC5384]

Attr_Type
0- Vector ;
4- Explicit RPF Vector ;

Other existing definitions are not related to RPF Vector Attribute.

The definition of Adjacency RPF Vector attribute

F bit: 0, indicating that the unrecognized device does not forward the attribute
E bit: indicates the last join attribute
Type: TBD
Length  depends on the address family of Encoded-Unicast address, including the length of 2 addresses.
Value Encoded-Unicast Address format defined in [RFC7761] Section 4.9.1, including 2 addresses. The first one indicates the address of the local interface, and the second one indicates the address of the peer interface. Only the case of the same address family is supported.

4.2. MLDP Label Mapping Message Extension

The LDP Label Mapping message format is defined in [RFC5036] Section 3.5.7. The MLDP P2MP protocol uses the message to establish a P2MP multicast tree. The Optional Parameters field can be extended to carry the node or link IP address list specified by the Segment List.

The TLV format definition in [RFC5036] Section 3.3 can be used for the Explicit Path TLV carrying the specified path of the Segment List.
The definition of Explicit Path TLV:

U bit  Unknown TLV bit. 1 indicates the unknown TLV MUST be silently ignored and the rest of the message processed as if the unknown TLV did not exist.

F bit  Forward unknown TLV bit. 0 indicates the unknown TLV is not forwarded with the containing message.

Type TBD1

Length contains all Sub-TLV lengths

Value Contains one or more Sub-TLVs, which are recorded in the order of TILFA’s Segment List. There are two types of Sub TLVs now. One of the two types is called Node Address Sub TLV which carries the node IP address corresponding to the NodeSID, and the other is called Adjacency Address Sub TLV which carries the local interface address and the peer interface address corresponding to the Adjacency SID.
Node Address Sub TLV carrying the node IP address corresponding to the NodeSID

U bit 1 indicates the unknown TLV MUST be silently ignored and the rest of the message processed as if the unknown TLV did not exist.

F bit 0 indicates the unknown TLV is not forwarded with the containing message.

E bit 1 indicates the last Sub TLV.

Type TBD2

Length IPv4 address 4 octet IPv6 address 16 octet

Value The IP address of the node corresponding to the NodeSID in the Segment List generated by TILFA

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1|0|E|      Type = TBD3        |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                                 |
|                 Local Interface Address                          |
|                 Remote Interface Address                         |
|                                                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Adjacency Address Sub TLV carrying the local interface address and the peer interface address corresponding to the Adjacency SID

U bit 1 indicates the unknown TLV MUST be silently ignored and the rest of the message processed as if the unknown TLV did not exist.

F bit 0 indicates the unknown TLV is not forwarded with the containing message.

E bit 1 indicates the last Sub TLV.

Type TBD3
Length IPv4 address 8 octet IPv6 address 32 octet

Value The IP address of the local interface and the IP address of the peer interface corresponding to the Adjacency SID in the Segment List generated by TILFA must be recorded in order, and MUST be the same address family.

5. IANA Considerations

This document requests IANA to assign a registry for Adjacency RPF Vector in PIM Join Attribute and the Explicit Path TLV Node Address Sub TLV, Adjacency Address Sub TLV in the Optional Parameters field of MLDP P2MP Label Mapping message. The assignment is requested permanent for IANA when this document is published as an RFC. The string TBD, TBD1, TBD2 and TBD3 should all be replaced by the assigned values accordingly.

6. Security Considerations

For general PIM-SM protocol Security Considerations, see [RFC7761].

For general MLDP protocol Security Considerations, see [RFC6388]

TBD

7. References

7.1. Normative References


7.2. Informative References

TBD

8. Acknowledgments

The authors would like to thank the following for their valuable contributions of this document:

TBD
Authors’ Addresses

Yisong Liu
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: liuyisong@huawei.com
Abstract

In an IGMP snooping multicast network with version 3 (v3) enabled on the routers, when a v2 join/leave is received for a multicast group the router operates on V2 compatible mode. For SSM ranges a (*,G)v2 or v3 report should be ignored by the router/switch. The IGMP snooping switches may not have knowledge about the user configured SSM range in the network to correctly discard/ignore the v2 join/leave. Accepting (*,G) v2 or v3 will cause SSM operations to fail. This draft discusses distribution of SSM ranges in the L2 multicast network so that L2 snooping switches can learn about the configured SSM ranges and discard any (*,G) v2/v3 reports for the said ranges.

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 January 9, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents
1.  Introduction

IGMP v2 join and leaves and IGMP v3 (*,G) group records should be discarded for Source specific multicast group ranges. The default SSM range is 232/8 but changing the range is possible. In a L2 multicast network the Snooping switches are unaware of the user configured SSM ranges in the network. Methods are needed to distribute user configured SSM ranges so that all snooping switches in the L2 domain knows about the same. Thus the snooping switches can discard the Version 2 joins/leaves falling in the SSM range. If the v2 joins/leaves for the SSM ranges are not discarded then the router/ querier start operating in v2 mode. This will result in outages. The same problem is applicable for MLD as well.

1.1.  Conventions Used in This Document

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].

1.2.  Terminology

DR:  Designated Router

SSM:  Source Specific Multicast
2. L2 network with a PIM router

In a LAN if a PIM router is detected the LAN segment should use the PIM SSM range configured on the PIM router which is the DR on the LAN. Snooping switches typically process PIM Hello packets already to detect routers. A new PIM Hello Option will carry the current (default or configured) SSM group ranges. The PIM Hello Option can be used by the snooping switches to learn the SSM ranges used in the network. Thus an IGMP message for a group in the SSM range in a v3 enabled network can correctly be discarded/ignored. Preventing hosts (whether by accident or a DoS attack) from disrupting the SSM service. Routers could be statically configured with the SSM group range. In case there are multiple routers on the LAN it is possible that routers are configured with different ranges. In that case, switches should use the range announced by the DR. The option allows for detecting configuration mistakes. A PIM router can log a message if it sees a neighbor announcing a different SSM range. Also, switches can log a message if they are statically configured with ranges that differ from what is announced by the DR. There is no hold time for the config. The config is removed if the router sends a hello without the option, or the DR expires. If a new DR is elected, the config will be replaced by what the new DR is announcing.

Figure 1: PIM SSM range hello option.

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Type = TBD           |      Length = Variable.       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Group Address 1 (Encoded-Group format)               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Group Address N (Encoded-Group format)               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

3. L2 multicast network with no PIM router

In a pure L2 only network a new IGMP message is sent from querier to learn the SSM ranges. The SSM range used should be configured on the querier and the querier will distribute it with a new message type so that all L2 switches can learn about the SSM range.
Figure 2: IGMP SSM range message.

<p>| | | | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| Type = TBD | Reserved | Checksum |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| Reserved | Num SSM ranges |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| SSM range 1 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Reserved |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| SSM range N |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

SSM range is an IP address plus a length octet.

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| SSM Prefix address |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| Prefix length |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

4. IANA Considerations

This document requires the assignment of a PIM hello option and an IGMP message type.

5. Acknowledgments

6. References

6.1. Normative References


Internet-DraSSM range distribution for L2 multicast networks. July 2019


6.2. Informative References


Authors' Addresses

Ramakrishnan Chokkanathapuram Sundaram
Cisco Systems, Inc.
Tasman Drive
San Jose CA 95134
USA

Email: ramaksun@cisco.com

Stig Venaas
Cisco Systems, Inc.
Tasman Drive
San Jose CA 95134
USA

Email: stig@cisco.com
Abstract

This document discusses the requirement of on-path telemetry for multicast traffic. The existing solutions are examined and their issues are addressed with new modifications that adapt to the multicast scenario.

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/.

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, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents
1. Introduction

Multicast traffic is an important traffic type in today’s Internet. Multicast provides services that are often real time (e.g., online meeting) or have strict QoS requirements (e.g., IPTV, Market Data). Multicast packet drop and delay can severely affect the application performance and user experience.

It is important to monitor the performance of the multicast traffic. Existing OAM techniques cannot gain direct and accurate information about the multicast traffic. New on-path telemetry techniques such as In-situ OAM [I-D.brockners-inband-oam-data] and Postcard-based Telemetry [I-D.song-ippm-postcard-based-telemetry] provide promising means to directly monitor the network experience of multicast traffic. However, multicast traffic has some unique characteristics which pose some challenges on efficiently applying such techniques.

This document describes the requirement for multicast traffic telemetry and shows the issues of the existing on-path telemetry techniques.
techniques. We then propose modifications to make these techniques adapt to the multicast application.

2. Requirements for Multicast Traffic Telemetry

Multicast traffic is forwarded through a multicast tree. With PIM and P2MP (MLDP, RSVP-TE) the forwarding tree is established and maintained by the multicast routing protocol. With BIER, no state is created in the network to establish a forwarding tree, instead, a bier header provides the necessary information for each packet to know the egress points. Multicast packets are only replicated at each tree branch node for efficiency.

There are several requirements for multicast traffic telemetry:

- Reconstrcut and visualize the multicast tree through data plane monitoring.
- Gather the multicast packet delay and jitter performance.
- Find the multicast packet drop location and reason.
- Gather the VPN state and tunnel information in case of P2MP multicast.

All of these requirements need to directly monitor the multicast traffic and derive data from the multicast packets. The conventional OAM mechanisms such as multicast ping and trace are insufficient to meet these requirements.

3. Issues of Existing Techniques

On-path Telemetry techniques that directly retrieve data from multicast traffic’s live network experience are ideal to address the above mentioned requirements. The representative techniques include In-situ OAM (IOAM) [I-D.brockners-inband-oam-data] and Postcard-based Telemetry (PBT) [I-D.song-ippm-postcard-based-telemetry]. However, unlike unicast, multicast poses some unique challenges to applying these techniques.

Multicast packets are replicated at each branch node in the corresponding multicast tree. Therefore, there are multiple copies of packets in the network.

If IOAM is used for on-path data collection, the partial trace data will also be replicated into multiple copies. The end result is that each copy of the multicast packet has a complete trace. Most of the data is redundant. Data redundancy introduces unnecessary header
overhead, wastes network bandwidth, and complicates the data processing. In case the multicast tree is large and the path is long, the redundancy problem becomes severe.

PBT can be used to eliminate such data redundancy, because each node on the tree only sends a postcard covering local data. However, PBT cannot track the tree branches properly so it can bring confusion about the multicast tree topology. For example, Node A has two branches, one to Node B and one to node D, and Node B leads to Node C and Node D leads to Node E. From the received postcard, one cannot tell whether or not Node C(E) is the next hop of Node B(D).

The fundamental reason for this problem is that there is not an identifier (either implicit or explicit) to correlate the data on each branch.

4. Proposed Modifications to Existing Techniques

We propose two solutions to address the above issues. One is built on PBT and requires augmentation to the instruction header of PBT-I; the other combines the IOAM trace mode and PBT for an optimized solution.

4.1. Per-hop postcard

The straightforward way to mitigate the PBT’s drawback is to augment it with a branch identifier field. Note that this only works for the PBT-I variation where an instruction header is present. To make the branch identifier globally unique, the branch node ID plus an index is used. For example, if Node A has two branches, to Nodes B and C, Node A will use \([A, 0]\) as the branch identifier for the branch to B, and \([A, 1]\) for the branch to C. The identifier is unchanged and carried with the multicast packet until the next branch node. Each postcard needs to include the branch identifier in the export data. The branch identifier and the other fields such as flow ID and sequence number are sufficient for the data analyzer to reconstruct the topology of the multicast tree.

Figure 1 shows an example of this solution. "P" stands for the postcard packet. The square brackets contain the branch identifier. The curly brace contains the telemetry data about a specific node.
4.2. Per-section postcard

The second solution is a combination of the IOAM trace mode and PBT. To avoid the data redundancy, at each branch node, the trace data accumulated so far is exported by a postcard before the packet is replicated. In this case, each branch still needs to keep some identifier to help correlate the postcards for each tree section. The natural way is to simply carry the branch node’s data (including its ID) in the trace of each branch. This is also necessary because each replicated multicast packet can have different telemetry data pertaining to this particular copy (e.g., node delay, egress timestamp, and egress interface). As a consequence of, the local data exported by each branch node can only contain partial data (e.g., ingress interface and ingress timestamp).

Figure 2 shows an example in a part of a multicast tree. Node B and D are two branch nodes so they will export a postcard covering the trace data for the previous section. The end node of each path need to export the data of the last section as a postcard as well.
There is no need to modify the IOAM trace mode header format. We just need to configure the branch node to export the postcard and refresh the IOAM header and data.

5. Considerations for Different Multicast Protocols

MTRACEv2 [RFC8487] provides an active probing approach for the tracing of an IP multicast routing path. Mtrace can also provide information such as the packet rates and losses, as well as other diagnostic information. New on-path telemetry techniques will enhance Mtrace, and other existing OAM solutions, with more granular and realtime network status data through direct measurements. There are various multicast protocols that are used to forward the multicast data. Each will require their own unique on-path telemetry solution.

5.1. Application in PIM

PIM-SM [RFC7761] is the most widely used multicast routing protocol deployed today. Of the various PIM modes (PIM-SM, PIM-DM, BIDIR-PIM, PIM-SSM), PIM-SSM is the preferred method due to its simplicity and removal of network source discovery complexity. With all PIM modes, control plane state is established in the network in order to forward multicast UDP data packets. But with PIM-SSM, the discovery of multicast sources is performed outside of the network via HTTP, SDN, etc. IP Multicast packets fall within the range of 224.0.0.0 through
239.255.255.255. The telemetry solution will need to work within this address range and provide telemetry data for this UDP traffic.

The proposed solutions for encapsulating the telemetry instruction header and metadata in IPv4/IPv6 UDP packets are described in [I-D.herbert-ipv4-udpencap-eh] and [I-D.ioametal-ippm-6man-ioam-ipv6-deployment].

5.2. Application in P2MP

Multicast Label Distribution Protocol (MLDP) and P2MP RSVP-TE are commonly used within a Multicast VPN (MVPN) environment. MLDP provides extensions to LDP to establish point-to-multipoint (P2MP) and multipoint-to-multipoint (MP2MP) label switched paths (LSPs) in MPLS networks. P2MP RSVP-TE provides extensions to RSVP-TE for establish traffic-engineered P2MP LSPs in MPLS networks. The telemetry solution will need to be able to follow these P2MP paths. The telemetry instruction header and data should be encapsulated into MPLS packets on P2MP paths. A corresponding proposal is described in [I-D.song-mpls-extension-header].

5.3. Application in BIER

BIER [RFC8279] adds a new header to multicast packets and allows the multicast packets to be forwarded according to the header only. By eliminating the requirement of maintaining per multicast group state, BIER is more scalable than the traditional multicast solutions.

OAM Requirements for BIER [I-D.ietf-bier-oam-requirements] lists many of the requirements for OAM at the BIER layer which will help in the forming of on-path telemetry requirements as well.

There is also current work to provide solutions for BIER forwarding in ipv6 networks. For instance, a solution, BIER in Non-MPLS IPv6 Networks [I-D.xie-bier-ipv6-encapsulation], proposes a new bier Option Type codepoint from the "Destination Options and Hop-by-Hop Options" IPv6 sub-registry. This is similar to what IOAM proposes for IPv6 transport.

Depending on how the BIER header is encapsulated into packets with different transport protocols, the method to encapsulate the telemetry instruction header and metadata also varies. It is also possible to make the instruction header and metadata a part of the BIER header itself, such as in a TLV.
6. Security Considerations

No new security issues are identified other than those discovered by the IOAM and PBT drafts.

7. IANA Considerations

The document makes no request of IANA.

8. Contributors

TBD

9. Acknowledgments

TBD

10. References

10.1. Normative References


10.2. Informative References

[I-D.brockners-inband-oam-data]

[I-D.herbert-ipv4-udpencap-eh]

[I-D.ietf-bier-oam-requirements]

[I-D.ioametal-ippm-6man-ioam-ipv6-deployment]

[I-D.song-ippm-postcard-based-telemetry]

[I-D.song-mpls-extension-header]
[I-D.xie-bier-ipv6-encapsulation]
Xie, J., Geng, L., McBride, M., Dhanaraj, S., Yan, G., and
Y. Xia, "Encapsulation for BIER in Non-MPLS IPv6
Networks", draft-xie-bier-ipv6-encapsulation-01 (work in
progress), June 2019.

Authors’ Addresses

Haoyu Song (editor)
Futurewei Technologies
2330 Central Expressway
Santa Clara
USA

Email: hsong@futurewei.com

Mike McBride
Futurewei Technologies
2330 Central Expressway
Santa Clara
USA

Email: mmcbride@futurewei.com
A Yang Data Model for IGMP/MLD Proxy
draft-zhao-pim-igmp-mld-proxy-yang-03.txt

Abstract

This document defines a YANG data model that can be used to configure and manage Internet Group Management Protocol (IGMP) or Multicast Listener Discovery (MLD) proxy devices. The YANG module in this document conforms to Network Management Datastore Architecture (NMDA).

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt
Table of Contents

1. Introduction...................................................3
   1.1. Terminology...............................................3
   1.2. Tree Diagrams.............................................3
2. Design of Data Model...........................................3
   2.1. Overview..................................................4
   2.2. Augment /rt:routing/rt:control-plane-protocols/rt:control-
   plane-protocol.................................................4
3. IGMP/MLD Proxy YANG Module.....................................5
4. Security Considerations.......................................13
5. IANA Considerations...........................................14
6. Normative References..........................................15
Authors’ Addresses...............................................17
1. Introduction

This document defines a YANG [RFC6020] data model for the management of Internet Group Management Protocol (IGMP) or Multicast Listener Discovery (MLD) proxy devices.

The YANG module in this document conforms to the Network Management Datastore Architecture defined in [RFC8342]. The "Network Management Datastore Architecture" (NMDA) adds the ability to inspect the current operational values for configuration, allowing clients to use identical paths for retrieving the configured values and the operational values.

1.1. Terminology

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].

The terminology for describing YANG data models is found in [RFC6020].

1.2. Tree Diagrams

A simplified graphical representation of the data model is used in this document. The meaning of the symbols in these diagrams is as follows:

- Brackets "[" and "]" enclose list keys.
- Abbreviations before data node names: "rw" means configuration (read-write), and "ro" means state data (read-only).
- Symbols after data node names: "?" means an optional node, "!" means a presence container, and "*" denotes a list and leaf-list.
- Parentheses enclose choice and case nodes, and case nodes are also marked with a colon (":")
- Ellipsis ("...") stands for contents of subtrees that are not shown.

2. Design of Data Model

The model covers Considerations for Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD) - Based Multicast Forwarding ("IGMP/MLD Proxying") [RFC4605].
The goal of this document is to define a data model that provides a common user interface to IGMP/MLD proxy. This document provides freedom for vendors to adapt this data model to their product implementations.

2.1. Overview

The IGMP/MLD proxy YANG module defined in this document has all the common building blocks for the IGMP/MLD proxy protocol.

The YANG module augments /rt:routing/rt:control-plane-protocols/rt:control-plane-protocol to enable IGMP/MLD proxy and configure other related parameters.

This YANG module follows the Guidelines for YANG Module Authors (NMDA) [draft-dsdt-nmda-guidelines-01]. This NMDA ("Network Management Datastore Architecture") architecture provides an architectural framework for datastores as they are used by network management protocols such as NETCONF [RFC6241], RESTCONF [RFC8040] and the YANG [RFC7950] data modeling language.

2.2. Augment /rt:routing/rt:control-plane-protocols/rt:control-plane-protocol

The YANG module augments /rt:routing/rt:control-plane-protocols/rt:control-plane-protocol to enable IGMP/MLD proxy under the upstream interface. There is also a constraint to make sure the upstream interface for IGMP/MLD proxy should not be configured PIM.

```
module: ietf-igmp-mld-proxy
augment /rt:routing/rt:control-plane-protocols/rt:control-plane-protocol:
  +--rw igmp-proxy {feature-igmp-proxy}?
    +--rw interfaces
      +--rw interface* [interface-name]
        +--rw interface-name if:interface-ref
        +--rw version? uint8
        +--rw enable? boolean
      +--ro group* [group-address]
        +--ro group-address inet:ipv4-address
        +--ro up-time? uint32
        +--ro filter-mode? enumeration
      +--ro source* [source-address]
        +--ro source-address inet:ipv4-address
        +--ro up-time? uint32
        +--ro filter-mode? enumeration
        +--ro downstream-interface* [interface-name]
          +--ro interface-name if:interface-ref
          +--ro filter-mode? enumeration
```
3. IGMP/MLD Proxy YANG Module

<CODE BEGINS> file ietf-igmp-mld-proxy@2019-07-03.yang
module ietf-igmp-mld-proxy {
  yang-version 1.1;
  // replace with IANA namespace when assigned
  prefix imp;

  import ietf-inet-types {
    prefix inet;
  }

  import ietf-interfaces {
    prefix if;
  }

  import ietf-routing {
    prefix rt;
  }

  import ietf-pim-base {
    prefix pim-base;
  }

  organization
    "IETF PIM Working Group";

...
The module defines a collection of YANG definitions common for all Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Proxy devices.

Copyright (c) 2019 IETF Trust and the persons identified as authors of the code. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, is permitted pursuant to, and subject to the license terms contained in, the Simplified BSD License set forth in Section 4.c of the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info).

This version of this YANG module is part of RFC XXXX; see the RFC itself for full legal notices.

revision 2019-07-03 {
  description  "Initial revision.";
  reference  "RFC XXXX: A YANG Data Model for IGMP and MLD Proxy";
}

/*
 * Features
 */
feature feature-igmp-proxy {
    description
        "Support IGMP Proxy protocol.";
    reference
        "RFC 4605";
}

feature feature-mld-proxy {
    description
        "Support MLD Proxy protocol.";
    reference
        "RFC 4605";
}

/*
 * Identities
 */

identity igmp-proxy {
    base rt:control-plane-protocol;
    description
        "IGMP Proxy protocol";
}

identity mld-proxy {
    base rt:control-plane-protocol;
    description
        "MLD Proxy protocol";
}

/*
 * Typedefs
 */

/*
 * Groupings
 */

grouping per-interface-config-attributes {
    description "Config attributes under interface view";
    leaf enable {
        type boolean;
        default false;
        description
            "Set the value to true to enable IGMP/MLD proxy";
    }
}
grouping state-group-attributes {
  description
    "State group attributes";

  leaf up-time {
    type uint32;
    units seconds;
    description
      "The elapsed time for (S,G) or (*,G).";
  }

  leaf filter-mode {
    type enumeration {
      enum "include" {
        description
          "In include mode, reception of packets sent to the specified multicast address is requested only from those IP source addresses listed in the source-list parameter";
      }
      enum "exclude" {
        description
          "In exclude mode, reception of packets sent to the given multicast address is requested from all IP source addresses except those listed in the source-list parameter.";
      }
      description
        "Filter mode for a multicast group, may be either include or exclude.";
    }
  }
}

/* augments */

augment "/rt:routing/rt:control-plane-protocols"+
    "/rt:control-plane-protocol" {

  description
    "IGMP Proxy augmentation to routing control plane protocol configuration and state."

  container igmp-proxy {
    when 'derived-from-or-self(/..rt:type, "imp:igmp-proxy")' {
      description
        "This container is only valid for IGMP Proxy protocol.";
    }
  }

if-feature feature-igmp-proxy;
description "IGMP proxy";
container interfaces {
  description
    "Containing a list of upstream interfaces.";

  list interface {
    key "interface-name";
    description
      "List of upstream interfaces.";

    leaf interface-name {
      type if:interface-ref;
      must "not( current() = /rt:routing"+
      "/rt:control-plane-protocols/pim-base:pim"+
      "/pim-base:interfaces/pim-base:interface"+
      "/pim-base:name )" {
        description
          "The upstream interface for IGMP proxy
          should not be configured PIM.";
      }
      description "The upstream interface name.";
    }

    leaf version {
      type uint8 {
        range "1..3";
      }
      default 2;
      description "IGMP version.";
    }

    uses per-interface-config-attributes;

  }

  list group {
    key "group-address";
    config false;
    description
      "Multicast group membership information
      that joined on the interface.";

    leaf group-address {
      type inet:ipv4-address;
      description
        "Multicast group address.";
    }

    uses state-group-attributes;
}
list source {
  key "source-address";
  description "List of multicast source information of the multicast group."
  leaf source-address {
    type inet:ipv4-address;
    description "Multicast source address"
  }
}

list downstream-interface {
  key "interface-name";
  description "The downstream interfaces list."
  leaf interface-name {
    type if:interface-ref;
    description "Downstream interfaces for each upstream-interface"
  }
  leaf filter-mode {
    type enumeration {
      enum "include" {
        description "In include mode, reception of packets sent to the specified multicast address is requested only from those IP source addresses listed in the source-list parameter";  
      }  
      enum "exclude" {
        description "In exclude mode, reception of packets sent to the given multicast address is requested from all IP source addresses except those listed in the source-list parameter";  
      }  
    }
    description "Filter mode for a multicast group, may be either include or exclude."
  }
}

} // list source
} // list group
} // interface
} // interfaces
}
augment "/rt:routing/rt:control-plane-protocols"+
"/rt:control-plane-protocol" {

description
"MLD Proxy augmentation to routing control plane protocol
configuration and state.";

container mld-proxy {
  when 'derived-from-or-self(../rt:type, "imp:mld-proxy")' {
    description
      "This container is only valid for MLD Proxy protocol.";
  }
  if-feature feature-mld-proxy;

description "MLD proxy";

container interfaces {
  description
    "Containing a list of upstream interfaces.";

  list interface {
    key "interface-name";
    description
      "List of upstream interfaces.";

    leaf interface-name {
      type if:interface-ref;
      must "not( current() = /rt:routing"+
        "/rt:control-plane-protocols/pim-base:pim"+
        "/pim-base:interfaces/pim-base:interface"+
        "/pim-base:name )" {
        description
          "The upstream interface for MLD proxy
          should not be configured PIM.";
      }
      description "The upstream interface name.";
    }

    leaf version {
      type uint8 {
        range "1..2";
      }
      default 2;
      description "MLD version.";
    }

    uses per-interface-config-attributes;
  }

  list group {
    key "group-address";
    config false;
    description
  }
}
"Multicast group membership information that joined on the interface."

leaf group-address {
  type inet:ipv6-address;
  description
    "Multicast group address.";
}

uses state-group-attributes;

list source {
  key "source-address";
  description
    "List of multicast source information of the multicast group.";
  leaf source-address {
    type inet:ipv6-address;
    description
      "Multicast source address";
  }

  uses state-group-attributes;

list downstream-interface {
  key "interface-name";
  description "The downstream interfaces list.";
  leaf interface-name {
    type if:interface-ref;
    description
      "Downstream interfaces for each upstream-interface";
  }

  leaf filter-mode {
    type enumeration {
      enum "include" {
        description
          "In include mode, reception of packets sent to the specified multicast address is requested only from those IP source addresses listed in the source-list parameter";
      }
      enum "exclude" {
        description
          "In exclude mode, reception of packets sent to the given multicast address is requested from all IP source addresses except those listed in the source-list parameter.";
      }
    }
  }
}

description
4. Security Considerations

The YANG module specified in this document defines a schema for data that is designed to be accessed via network management protocols such as NETCONF [RFC6241] or RESTCONF [RFC8040]. The lowest NETCONF layer is the secure transport layer, and the mandatory-to-implement secure transport is Secure Shell (SSH) [RFC6242]. The lowest RESTCONF layer is HTTPS, and the mandatory-to-implement secure transport is TLS [RFC5246].

The NETCONF access control model [RFC6536] provides the means to restrict access for particular NETCONF or RESTCONF users to a preconfigured subset of all available NETCONF or RESTCONF protocol operations and content.

There are a number of data nodes defined in this YANG module that are writable/creatable/deletable (i.e., config true, which is the default). These data nodes may be considered sensitive or vulnerable in some network environments. Write operations (e.g., edit-config) to these data nodes without proper protection can have a negative effect on network operations. These are the subtrees and data nodes and their sensitivity/vulnerability:

/rt:routing/rt:control-plane-protocols/rt:control-plane-protocol

Unauthorized access to any data node of these subtrees can adversely affect the IGMP/MLD proxy subsystem of both the local device and the network. This may lead to network malfunctions, delivery of packets to inappropriate destinations, and other problems.

Some of the readable data nodes in this YANG module may be considered sensitive or vulnerable in some network environments. It is thus...
important to control read access (e.g., via get, get-config, or notification) to these data nodes. These are the subtrees and data nodes and their sensitivity/vulnerability:

/rt:routing/rt:control-plane-protocols/rt:control-plane-protocol

Unauthorized access to any data node of these subtrees can disclose the operational state information of IGMP/MLD proxy on this device.

5. IANA Considerations

RFC Ed.: In this section, replace all occurrences of 'XXXX' with the actual RFC number (and remove this note).

This document registers the following namespace URIs in the IETF XML registry [RFC3688]:

```
Registrant Contact: The IESG.
XML: N/A, the requested URI is an XML namespace.
```

This document registers the following YANG modules in the YANG Module Names registry [RFC7950]:

```
name:        ietf-igmp-mld-proxy
prefix:      imp
reference:   RFC XXXX
```

Zhao & Liu, etc Expires January 02, 2020 [Page 14]
6. Normative References


Authors’ Addresses

Hongji Zhao
Ericsson (China) Communications Company Ltd.
Ericsson Tower, No. 5 Lize East Street,
Chaoyang District Beijing 100102, P.R. China
Email: hongji.zhao@ericsson.com

Xufeng Liu
Volta Networks
USA
EMail: Xufeng.liu.ietf@gmail.com

Yisong Liu
Huawei Technologies
Huawei Bld., No.156 Beijing Rd.
Beijing 100095
China
Email: liuyisong@huawei.com

Mani Panchanathan
Cisco
India
Email: mapancha@cisco.com

Mahesh Sivakumar
Juniper Networks
1133 Innovation Way
Sunnyvale, California
USA
EMail: sivakumar.mahesh@gmail.com