

PIM Working Group
Internet-Draft
Intended status: Standards Track
Expires: 12 June 2022

G. Mirsky
Ericsson
J. Xiaoli
ZTE Corporation
9 December 2021

Fast Failover in Protocol Independent Multicast - Sparse Mode (PIM-SM)
Using Bidirectional Forwarding Detection (BFD) for Multipoint Networks
draft-ietf-pim-bfd-p2mp-use-case-10

Abstract

This document specifies how Bidirectional Forwarding Detection for multipoint networks can provide sub-second failover for routers that participate in Protocol Independent Multicast - Sparse Mode (PIM-SM). An extension to the PIM Hello message used to bootstrap a point-to-multipoint BFD session is 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 12 June 2022.

Copyright Notice

Copyright (c) 2021 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 include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
1.1. Conventions used in this document	3
1.1.1. Terminology	3
1.1.2. Requirements Language	3
2. BFD Discriminator PIM Hello Option	3
2.1. Using P2MP BFD in PIM Router Monitoring	4
2.2. P2MP BFD in PIM DR Load Balancing	5
2.3. Multipoint BFD Encapsulation	5
3. IANA Considerations	6
4. Security Considerations	6
5. Acknowledgments	6
6. References	6
6.1. Normative References	6
6.2. Informative References	7
Authors' Addresses	7

1. Introduction

Faster convergence in the control plane minimizes the periods of traffic blackholing, transient routing loops, and other situations that may negatively affect service data flow. Faster convergence in the control plane is beneficial 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. A conforming implementation of PIM-SM elects a Designated Router (DR) on each PIM-SM interface. When a group of PIM-SM nodes is connected to a shared media segment, e.g., Ethernet, the node elected as DR acts 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.

Bidirectional Forwarding Detection (BFD) [RFC5880] was originally defined to detect a failure of a point-to-point (p2p) path, single-hop [RFC5881] or multihop [RFC5883]. In some PIM-SM deployments, a

p2p BFD can be used to detect a failure and enable faster failover. [RFC8562] extends the BFD base specification [RFC5880] for multipoint and multicast networks, which matches the deployment scenarios for PIM-SM over a LAN segment. A BFD system in p2mp environment that transmits BFD Control messages using the BFD Demand mode [RFC5880] creates less BFD state than the Asynchronous mode. Point-to-multipoint (p2mp) BFD can enable faster detection of PIM-SM router failure compared to PIM-SM without BFD and thus minimize multicast service disruption. The monitored PIM-SM router acts as the head and other routers as tails of a p2mp BFD session. This document defines the monitoring of a PIM-SM router using p2mp BFD. This document also defines the extension to PIM-SM [RFC7761] to bootstrap a PIM-SM router to join in p2mp BFD session over shared media segment.

1.1. Conventions used in this document

1.1.1. Terminology

This document uses terminology defined in [RFC5880], [RFC8562], and [RFC7761]. Familiarity with these specifications and the terminology used is expected.

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. BFD Discriminator PIM Hello Option

Figure 1 displays the new optional BFD Discriminator PIM Hello option to bootstrap a tail of the p2mp BFD session.

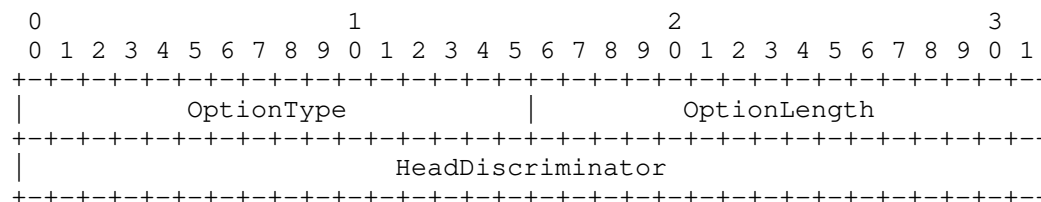


Figure 1: BFD Discriminator PIM Hello Option

where new fields are interpreted as:

OptionType: TBA.

OptionLength: MUST be set to 4.

HeadDiscriminator: the four-octet field MUST be included in the BFD Discriminator PIM-SM Hello option. The value MUST NOT be zero. It equals the value of My Discriminator ([RFC5880]) allocated by the head.

If the value of the OptionLength field is not equal to 4, the BFD Discriminator PIM Hello option is considered malformed, and the receiver MUST stop processing PIM Hello options. If the value of the HeadDiscriminator field equals zero, then the BFD Discriminator PIM Hello option MUST be considered invalid, and the receiver MUST ignore it. The receiver SHOULD log a notification regarding the malformed or invalid BFD Discriminator Hello option under the control of a throttling logging mechanism.

2.1. Using P2MP BFD in PIM Router Monitoring

If the head is no longer serving the function that prompted it to be monitored, then it MUST cease including the BFD Discriminator PIM Hello option in its PIM-Hello message, and it SHOULD shut down the BFD session following the procedures described in Section 5.9 [RFC8562].

The head MUST create a BFD session of type MultipointHead [RFC8562]. Note that any PIM-SM router, regardless of its role, MAY become a head of a p2mp BFD session. To control the volume of BFD control traffic on a shared media segment, an operator should carefully select PIM-SM routers configured as a head of a p2mp BFD session. The head MUST include the BFD Discriminator PIM Hello option in its PIM Hello messages.

A PIM-SM router that is configured to monitor the head by using p2mp BFD is referred to throughout this document as a "tail". When such a tail receives a PIM-Hello packet with the BFD Discriminator PIM Hello option, the tail MAY create a p2mp BFD session of type MultipointTail, as defined in [RFC8562].

The node that includes the BFD Discriminator PIM Hello option transmits BFD Control packets periodically. For the tail to correctly demultiplex BFD [RFC8562], the source address and My Discriminator of the BFD packets MUST be the same as the source address and the HeadDiscriminator, respectively, of the PIM Hello message. If that is not the case, the tail BFD node would not be able to monitor the state of the PIM-SM node, that is, the head of the p2mp BFD session, though the regular PIM-SM mechanisms remain fully operational.

If the tail detects a MultipointHead failure [RFC8562], it MUST delete the corresponding neighbor state and follow procedures defined in [RFC7761] for the DR and additional neighbor state deletion after the neighbor timeout expires.

If the head ceases to include the BFD Discriminator PIM Hello option in its PIM-Hello message, tails SHOULD close the corresponding MultipointTail BFD session without affecting the PIM state in any way. Thus, the tail stops using BFD to monitor the head and reverts to the procedures defined in [RFC7761].

2.2. P2MP BFD in PIM DR Load Balancing

[RFC8775] specifies the PIM Designated Router Load Balancing (DRLB) functionality. Any PIM router that advertises the DRLB-Cap Hello Option can become the head of a p2mp BFD session, as specified in Section 2.1. The head router administratively sets the bfd.SessionState to Up in the MultipointHead session [RFC8562] only if it is a Group Designated Router (GDR) Candidate, as specified in Sections 5.5 and 5.6 of [RFC8775]. If the router is no longer the GDR, then it MUST shut down following the procedures described in Section 5.9 [RFC8562]. For each GDR Candidate that includes BFD Discriminator option in its PIM Hello, the PIM DR MUST create a MultipointTail session [RFC8562]. PIM DR demultiplexes BFD sessions based on the value of the My Discriminator field and the source IP address. If PIM DR detects a failure of one of the sessions, it MUST remove that router from the GDR Candidate list and immediately transmit a new DRLB-List option.

2.3. Multipoint BFD Encapsulation

The MultipointHead of a p2mp BFD session when transmitting BFD Control packets:

MUST set TTL or Hop Limit value to 255 (Section 5 [RFC5881]). Similarly, all received BFD Control packets that are demultiplexed to the session MUST be discarded if the received TTL or Hop Limit is not equal to 255;

MUST use the group address ALL-PIM-ROUTERS ('224.0.0.13' for IPv4 and 'ff02::d' for IPv6) as destination IP address

3. IANA Considerations

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

Value	Length	Name	Reference
TBA	4	BFD Discriminator Option	This document

Table 1: BFD Discriminator option type

4. Security Considerations

This document defines a way to accelerate detecting a failure that affects PIM functionality by using BFD. The operation of either protocol is not changed.

The security considerations discussed in [RFC7761], [RFC5880], [RFC5881], [RFC8562], and [RFC8775] apply to this document.

5. Acknowledgments

The authors cannot say enough to express their appreciation of the comments and suggestions we received from Stig Venaas. The authors greatly appreciate the comments and suggestions by Alvaro Retana that improved the clarity of this document.

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.

- [RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, DOI 10.17487/RFC5881, June 2010, <<https://www.rfc-editor.org/info/rfc5881>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <<https://www.rfc-editor.org/info/rfc7761>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8562] Katz, D., Ward, D., Pallagatti, S., Ed., and G. Mirsky, Ed., "Bidirectional Forwarding Detection (BFD) for Multipoint Networks", RFC 8562, DOI 10.17487/RFC8562, April 2019, <<https://www.rfc-editor.org/info/rfc8562>>.
- [RFC8775] Cai, Y., Ou, H., Vallepalli, S., Mishra, M., Venaas, S., and A. Green, "PIM Designated Router Load Balancing", RFC 8775, DOI 10.17487/RFC8775, April 2020, <<https://www.rfc-editor.org/info/rfc8775>>.

6.2. Informative References

- [RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, DOI 10.17487/RFC5883, June 2010, <<https://www.rfc-editor.org/info/rfc5883>>.

Authors' Addresses

Greg Mirsky
Ericsson

Email: gregimirsky@gmail.com

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

Email: ji.xiaoli@zte.com.cn

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 11 May 2022

V. Kamath
VMware
R. Chokkanathapuram Sundaram
Cisco Systems, Inc.
R. Banthia
Apstra
A. Gopal
Cisco Systems, Inc.
7 November 2021

PIM Null-Register packing
draft-ietf-pim-null-register-packing-11

Abstract

In PIM-SM networks PIM Null-Register messages are sent by the Designated Router (DR) to the Rendezvous Point (RP) to signal the presence of Multicast sources in the network. There are periodic PIM Null-Registers sent from the DR to the RP to keep the state alive at the RP as long as the source is active. The PIM Null-Register message 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 Packed Null-Register message. We will refer to the new packed formats as the PIM Packed Null-Register format and PIM Packed Register-Stop format throughout the document. This document also discusses interoperability between the PIM routers which do not understand the PIM Packed Null-Register format and routers which do understand it.

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 11 May 2022.

Copyright Notice

Copyright (c) 2021 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 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.

Table of Contents

1. Introduction	2
1.1. Conventions used in this document	3
1.2. Terminology	3
2. Packed Null-Register Capability	3
3. PIM Packed Null-Register message format	4
4. PIM Packed Register-Stop message format	5
5. Protocol operation	6
6. Operational Considerations	7
7. PIM Anycast RP Considerations	7
8. PIM RP router version downgrade	7
9. Fragmentation Considerations	7
10. Security Considerations	8
11. IANA Considerations	8
12. Acknowledgments	8
13. References	8
13.1. Normative References	8
13.2. Informative References	9
Authors' Addresses	9

1. Introduction

PIM Null-Registers are sent by the DR 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 messages that are sent also increases. This results in more PIM packet processing at the RP and the DR.

The control plane policing (COPP), monitors the packets that are processed by the control plane. The high rate at which Null-Registers are received at the RP can lead to COPP drops of Multicast PIM Null-Register messages. This draft proposes a method to efficiently pack multiple PIM Null-Registers [RFC7761] (Section 4.4) and Register-Stops [RFC7761] (Section 3.2) into a single message as these packets anyway do not contain encapsulated data.

The draft also discusses interoperability with PIM 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", "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

RP: Rendezvous Point

DR: Designated Router

2. Packed Null-Register Capability

A router (DR) can decide to pack multiple Null-Register messages based on the capability received from the RP as part of the PIM Register-Stop. This ensures compatibility with routers that do not support processing of the new format. The capability information can be indicated by the RP via the PIM Register-Stop message sent to the DR. Thus a DR will switch to the new format only when it learns that the RP is capable of handling the PIM Packed Null-Register messages.

Conversely, a DR that does not support the packed format can continue generating the PIM Null-Register as defined in [RFC7761] (Section 4.4). 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.

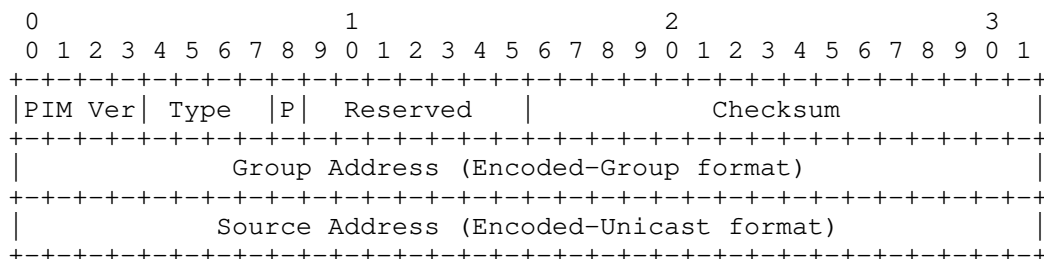


Figure 1: PIM Register-Stop message with capability option

PIM Version, Type, Checksum, Group Address, Source Address:

Same as [RFC7761] (Section 4.9.4)

P:

Capability bit (flag bit 7) used to indicate support for the
Packed Null-Register Capability

3. PIM Packed Null-Register message format

PIM Packed Null-Register message format includes a count to indicate the number of Null-Register records in the message.

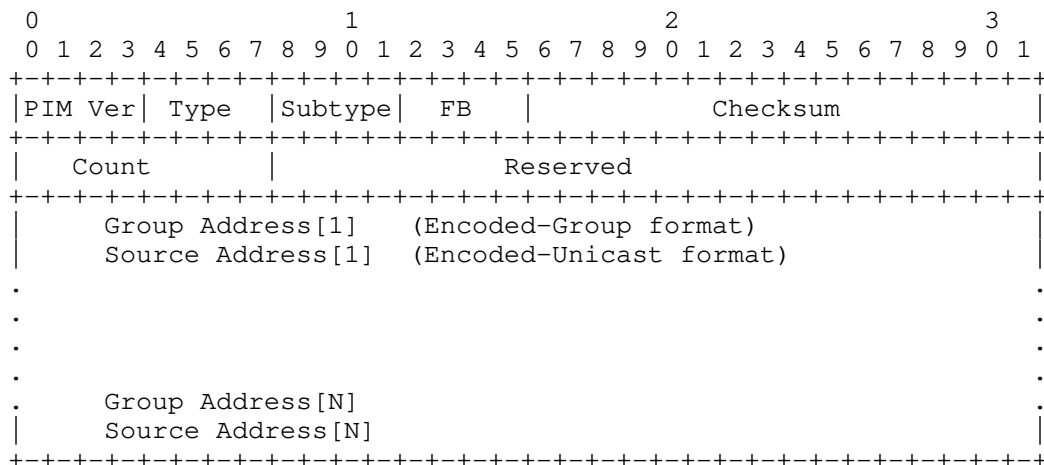


Figure 2: PIM Packed Null-Register message format

PIM Version, Reserved, Checksum:

Same as [RFC7761] (Section 4.9.3)

Type, SubType:

The new packed Null-Register Type and SubType values TBD.
[RFC8736]

Count:

The number of packed Null-Register records. A record consists of a Group Address and Source Address pair.

Group Address, Source Address:

Same as [RFC7761] (Section 4.9.4)

4. PIM Packed Register-Stop message format

The PIM Packed Register-Stop message includes a count to indicate the number of records that are present in the message.

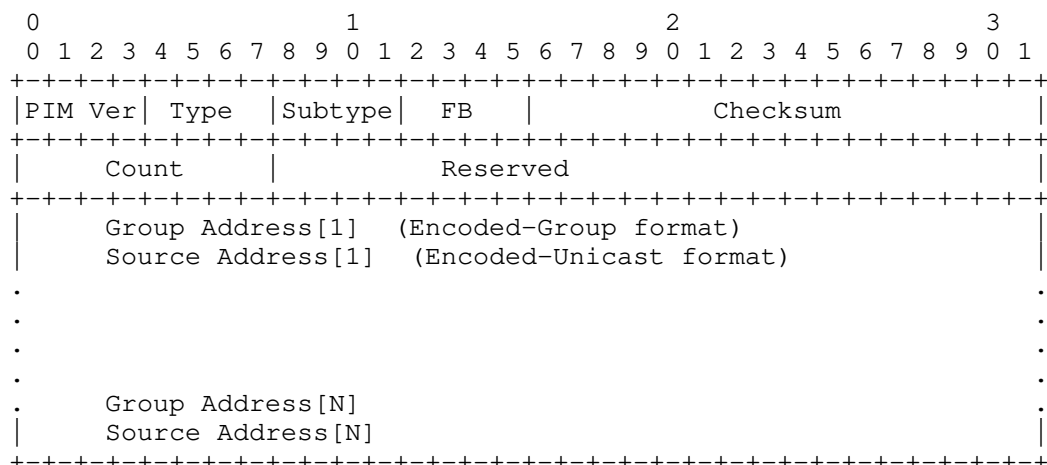


Figure 3: PIM Packed Register-Stop message format

PIM Version, Reserved, Checksum:

Same as [RFC7761] (Section 4.9.4)

Type:

The new Register Stop Type and SubType values TBD

Count:

The number of PIM packed Register-Stop records. A record consists of a Group Address and Source Address pair.

Group Address, Source Address:

Same as [RFC7761] (Section 4.9.4)

5. Protocol operation

The following combinations exist -

1. DR and RP both support the PIM Packed Null-Register and PIM Packed Register-Stop formats:
 - * As specified in [RFC7761], the DR sends PIM Register messages towards the RP when a new source is detected.
 - * An RP supporting this specification MUST set the P-bit in the corresponding Register-Stop messages.
 - * When a Register-Stop message with the P-bit set is received, the DR SHOULD send PIM Packed Null-Register messages (Section 3) to the RP instead of multiple Register messages with the N-bit set [RFC7761].
 - * The RP, after receiving a PIM Packed Null-Register message SHOULD start sending PIM Packed Register-Stop messages (Section 4) to the corresponding DR instead of individual Register-Stop messages.
2. DR supports but RP does not support the PIM Packed Null-Register and PIM Packed Register-Stop formats:
 - * As specified in [RFC7761], DR sends PIM Null-Registers towards the RP.
 - * After receiving DR's PIM Null-Register message, RP sends a normal Register-Stop without any capability information.
 - * DR then sends PIM Null-Registers in the unpacked format [RFC7761].
3. RP supports but DR does not support the PIM Packed Null-Register and PIM Packed Register-Stop formats:

- * As specified in [RFC7761], DR sends the PIM Null-Register towards the RP.
- * After receiving DR's PIM Null-Register message, RP sends a PIM Packed Register-Stop towards the DR that includes capability information.
- * Since DR does not support the new format, it sends PIM Null-Registers in the unpacked format [RFC7761].

6. Operational Considerations

In case the network manager disables the packed capability at the RP, the router should not advertise the capability. However, an implementation MAY choose to still parse any packed registers if they are received. This may be particularly useful in the transitional period after the network manager disables it.

7. PIM Anycast RP Considerations

The PIM Packed Null-Register format should be enabled only if it is supported by all PIM Anycast RP [RFC4610] members in the RP set for the RP address. This consideration applies to PIM Anycast RP with MSDP [RFC3446] as well.

8. PIM RP router version downgrade

Consider a PIM RP router that supports PIM Packed Null-Registers and PIM Packed Register-Stops. When this router downgrades to a software version which does not support PIM Packed Null-Registers and PIM Packed Register-Stops, the DR that sends the PIM Packed Null-Register message will not get a PIM Register-Stop message back from the RP. In such scenarios the DR can send an unpacked PIM Null-Register and check the PIM Register-Stop to see if the capability bit (P-bit) for PIM Packed Null-Register is set or not. If it is not set then the DR will continue sending unpacked PIM Null-Register messages.

9. Fragmentation Considerations

When building a PIM Packed Null-Register message or PIM Packed Register-Stop message, a router should include as many records as possible based on the path MTU towards RP, if path MTU discovery is done. Otherwise, the number of records should be limited by the MTU of the outgoing interface.

10. Security Considerations

General Register messages security considerations from [RFC7761] apply. As mentioned in [RFC7761], PIM Null-Register messages and Register-Stop messages are forwarded by intermediate routers to their destination using normal IP forwarding. Without data origin authentication, an attacker who is located anywhere in the network may be able to forge a Null-Register or Register-Stop message. We next consider the effect of a forgery of each of these messages. By forging a Register message, an attacker can cause the RP to inject forged traffic onto the shared multicast tree.

By forging a Register-Stop message, an attacker can prevent a legitimate DR from registering packets to the RP. This can prevent local hosts on that LAN from sending multicast packets. The above two PIM messages are not changed by intermediate routers and need only be examined by the intended receiver. Thus, these messages can be authenticated end-to-end. Attacks on Register and Register-Stop messages do not apply to a PIM-SSM-only implementation, as these messages are not used in PIM-SSM.

There is another case where a spoofed Register-Stop can be sent to make it appear that is from the RP, and that the RP supports this new packed capability when it does not. This can cause Null-Registers to be sent to an RP that doesn't support this packed format. But standard methods to prevent spoofing should take care of this case. For example, uRPF can be used to filter out packets coming from the outside from addresses that belong to routers inside.

11. IANA Considerations

This document requires the assignment of Capability bit (P-bit), flag bit 7 in the PIM Register-Stop message.

This document requires the assignment of 2 new PIM message types for the PIM Packed Null-Register and PIM Packed Register-Stop.

12. Acknowledgments

The authors would like to thank Stig Venaas, Anish Peter, Zheng Zhang and Umesh Dudani for their helpful comments on the draft.

13. References

13.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <<https://www.rfc-editor.org/info/rfc7761>>.
- [RFC4610] Farinacci, D. and Y. Cai, "Anycast-RP Using Protocol Independent Multicast (PIM)", RFC 4610, DOI 10.17487/RFC4610, August 2006, <<https://www.rfc-editor.org/info/rfc4610>>.
- [RFC8736] Venaas, S. and A. Retana, "PIM Message Type Space Extension and Reserved Bits", RFC 8736, DOI 10.17487/RFC8736, February 2020, <<https://www.rfc-editor.org/info/rfc8736>>.

13.2. Informative References

- [RFC3446] Kim, D., Meyer, D., Kilmer, H., and D. Farinacci, "Anycast Rendezvous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)", RFC 3446, DOI 10.17487/RFC3446, January 2003, <<https://www.rfc-editor.org/info/rfc3446>>.

Authors' Addresses

Vikas Ramesh Kamath
VMware
3401 Hillview Ave
Palo Alto, CA 94304
United States of America

Email: vkamath@vmware.com

Ramakrishnan Chokkanathapuram Sundaram
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
United States of America

Email: ramaksun@cisco.com

Raunak Banthia
Apstra
333 Middlefield Rd STE 200
Menlo Park, CA 94025
United States of America

Email: rbanthia@apstra.com

Ananya Gopal
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
United States of America

Email: ananygop@cisco.com

PIM Working Group
Internet Draft
Intended status: Informational
Expires: September 4, 2022

Yisong Liu
China Mobile
M. McBride
Futurewei
Z. Zhang
ZTE
J. Xie
Huawei
C. Lin
New H3C Technologies
Mar 4, 2022

Multicast-only Fast Reroute Based on Topology Independent Loop-free
Alternate Fast Reroute
draft-liu-pim-mofrr-tilfa-05

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/lid-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 September 4, 2022.

Copyright Notice

Copyright (c) 2022 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
4. IANA Considerations.....	8
5. Security Considerations.....	8
6. References.....	8
6.1. Normative References.....	8
6.2. Informative References.....	9
7. Acknowledgments.....	9
Authors' Addresses.....	10

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). 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 use the terms defined in [RFC7431], and also use 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 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 cannot establish a standby multicast tree and cannot implement MoFRR protection. Therefore, the current MoFRR of PIM 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 P space node. PIM 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.

PIM join 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 protocol packets are transmitted only in unicast mode, then equivalently the PIM 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 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. So it can form an end-to-end complete multicast tree for forwarding multicast traffic. Therefore, it is not possible to send PIM packets like unicast traffic according to the Segment List path and can only establish a standby multicast tree.

3. Solution

A secondary 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 the Segment List calculated by TILFA.

It is available in principle that the path information of the Segment List is added to the PIM 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.

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.

This document can use the above two RPF Vector standards and does not need to extend the PIM 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.

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 can be looked as the local address of the incoming interface, and IP-Lb can be looked as the address of the upstream neighbor. So IP-Lb can be added to the PIM join packet as the explicit RPF Vector attribute.

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 the PIM upstream neighbor. The node A can find the incoming interface in the unicast routing table according to the IP-Lb and IP-Lb is used as the RPF upstream address of the PIM join packet to the node B.

After the PIM join packet is received on the node B, the PIM protocol can find no more RPF Vector 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.

4. Illustration

This section provides an illustration of MoFRR based on TI-LFA. The example topology is shown in Figure 1. The metric of R3-R4 link is 100, and the metrics of other links are 10. All the link metrics are bidirectional.

```

<-----Priamry Path--- (S,G) Join
[S]----(R1)---(R2)***** (R6)-----[R]
      |           |         |
      <--          |         |
              (R5)    |
                  |

```

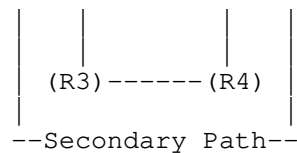


Figure 1: Sample Topology

The IP addresses and MPLS SIDs, which may be involved in the MoFRR calculation, are configured as following:

Node	IP Address	Node SID
R4	4.4.4.4/32	16004

Link	IP Address	Adjacency SID
R3->R4	14.0.0.1/24	24001
R4->R3	14.0.0.2/24	24002

The primary path of the PIM join packet is R6->R2->R1, and the secondary path is R6->R5->R4->R3->R2->R1. However, the traditional LFA does not work properly for the secondary path, because the shortest path to R2 from R5 (or from R4) still goes through R6-R2 link. In this case, R6 needs to calculate the secondary UMH using the proposed MoFRR method based on TI-LFA.

According to the TI-LFA algorithm, P-Space and Q-Space are shown in Figure 2. The TI-LFA repair path consists of the Node SID of R4 and the Adjacency SID of R4->R3. The repair segment list is (16004, 24002).

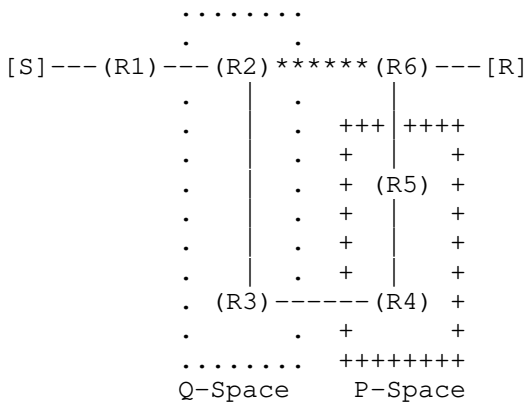


Figure 2: P-Space and Q-Space

In the procedure of PIM, the IP addresses associated with the repair segment list are looked up in the IGP link state database. The Node SID 16004 corresponds to 4.4.4.4, which will be carried in the RPF Vector Attribute. The Adjacency SID 24002 corresponds to local address 14.0.0.2 and remote peer address 14.0.0.1, and 14.0.0.1 will be carried in the Explicit RPF Vector Attribute. Therefore, R6 installs the secondary UMH with these RPF Vectors.

The forwarding of PIM join packet along the secondary path is shown in Figure 3.

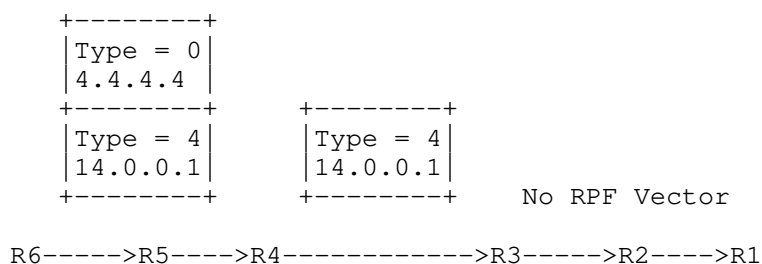


Figure 3: Forwarding PIM Join Packet along Secondary Path

R6 inserts two RPF Vector Attributes in the PIM join packet, which are 4.4.4.4 of Type 0 (RPF Vector Attribute) and 14.0.0.1 of Type 4 (Explicit RPF Vector Attribute). Then R6 forwards the packet along the secondary path.

When R5 receives the packet, R5 performs a unicast route lookup of the first RPF Vector 4.4.4.4 and sends the packet to R4.

R4 is the owner of 4.4.4.4, so it removes the first RPF Vector from the packet and forwards according to the following RPF Vector. R4 sends the packet to R3 according to the next RPF Vector 14.0.0.1, since its PIM neighbor R3 corresponds to 14.0.0.1.

When R3 receives the packet, as the owner of 14.0.0.1, it removes the RPF Vector. Then the packet has no RPF Vector, and will be forwarded to the source through R3->R2->R1 according to unicast routes.

After the PIM join packet reaches R1, a secondary multicast tree, R1->R2->R3->R4->R5->R6, is established hop-by-hop for (S, G) by MoFRR based on TI-LFA.

The above procedures can also work in IPv6 data plane. The TI-LFA path computation algorithm in the SRv6 data plane is the same as in the SR-MPLS data plane. Instead of MPLS labels, SRv6 SIDs are used

to build repair list. Similarly, the RPF Vector Attributes and the Explicit RPF Vector Attributes will contain IPv6 addresses instead of IPv4 addresses.

5. IANA Considerations

No IANA actions are required for this document.

6. Security Considerations

This document does not change the security properties of PIM. For general PIM-SM protocol Security Considerations, see [RFC7761].

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5384] Boers, A., Wijnands, I., and E. Rosen, "The Protocol Independent Multicast (PIM) Join Attribute Format", RFC 5384, November 2008
- [RFC5496] Wijnands, IJ., Boers, A., and E. Rosen, "The Reverse Path Forwarding (RPF) Vector TLV", RFC 5496, March 2009
- [RFC7431] Karan, A., Filsfils, C., Wijnands, IJ., Ed., and B. Decraene, "Multicast-Only Fast Reroute", RFC 7431, August 2015
- [RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", RFC 7490, April 2015
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 7761, March 2016
- [RFC7891] Asghar, J., Wijnands, IJ., Ed., Krishnaswamy, S., Karan, A., and V. Arya, "Explicit Reverse Path Forwarding (RPF) Vector", RFC 7891, June 2016
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, May 2017

[I-D.ietf-rtgwg-segment-routing-ti-lfa] Litkowski, S., Bashandy, A., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", draft-ietf-rtgwg-segment-routing-ti-lfa-08, work-in-progress, January 2022

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
China Mobile
China
Email: liuyisong@chinamobile.com

Mike McBride
Futurewei Inc.
USA
Email: michael.mcbride@futurewei.com

Zheng (Sandy) Zhang
ZTE Corporation
China
Email: zzhang_ietf@hotmail.com

Jingrong Xie
Huawei Technologies
China
Email: xiejingrong@huawei.com

Changwang Lin
New H3C Technologies
China
Email: linchangwang.04414@h3c.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 9, 2020

R. Chokkanathapuram Sundaram
S. Venaas
Cisco Systems, Inc.
July 8, 2019

Source specific multicast range distribution for L2 multicast networks
draft-ramki-igmp-ssm-ranges-00

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

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.

Table of Contents

1. Introduction	2
1.1. Conventions Used in This Document	2
1.2. Terminology	2
2. L2 network with a PIM router	3
3. L2 multicast network with no PIM router	3
4. IANA Considerations	4
5. Acknowledgments	4
6. References	4
6.1. Normative References	4
6.2. Informative References	5
Authors' Addresses	5

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

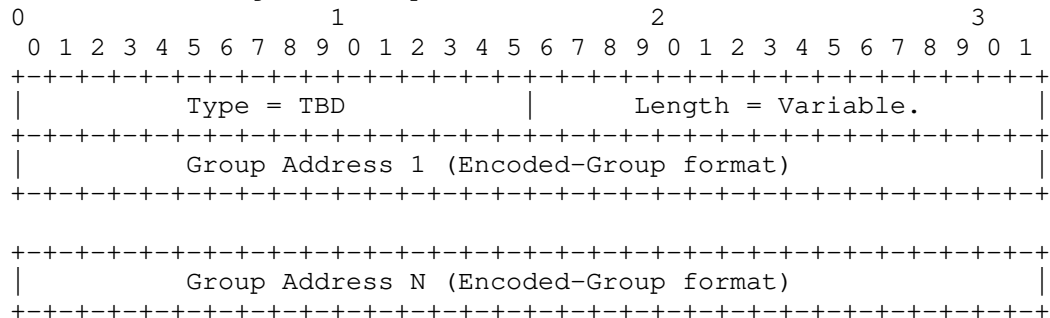
DR: Designated Router

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 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.

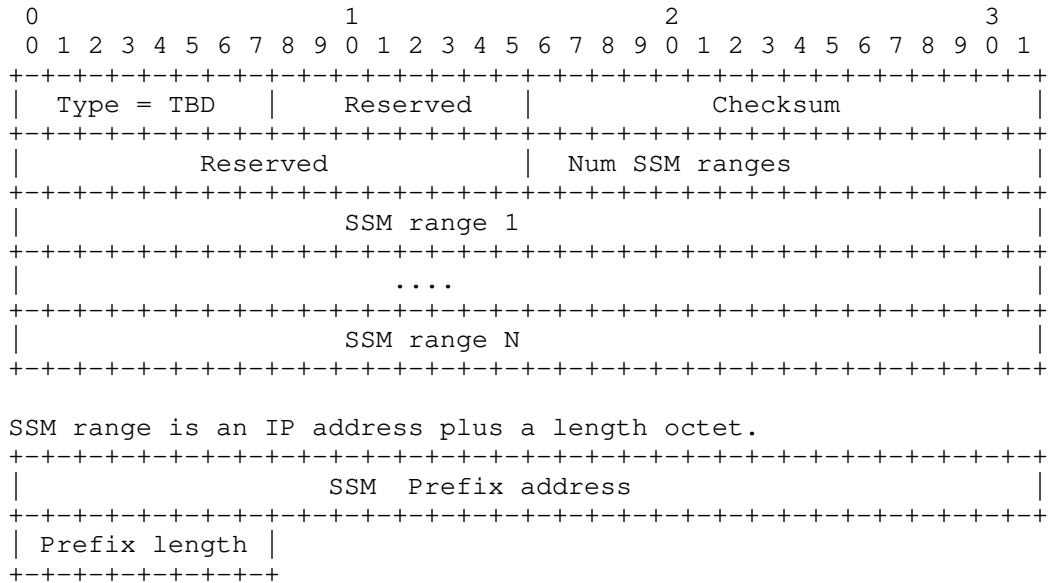
Figure 1: PIM SSM range hello option.



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.



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

- [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>>.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <<https://www.rfc-editor.org/info/rfc3376>>.
- [RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, DOI 10.17487/RFC4604, August 2006, <<https://www.rfc-editor.org/info/rfc4604>>.

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

[RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, DOI 10.17487/RFC4607, August 2006, <<https://www.rfc-editor.org/info/rfc4607>>.

6.2. Informative References

[RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, DOI 10.17487/RFC3973, January 2005, <<https://www.rfc-editor.org/info/rfc3973>>.

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

MBONED
Internet-Draft
Intended status: Standards Track
Expires: July 23, 2021

H. Song
M. McBride
Futurewei Technologies
G. Mirsky
ZTE Corp.
G. Mishra
Verizon Inc.
January 19, 2021

Multicast On-path Telemetry Solutions
draft-song-multicast-telemetry-07

Abstract

This document discusses the requirement of on-path telemetry for multicast traffic. The existing solutions are examined and their issues are identified. Solution modifications are proposed to allow the original multicast tree to be correctly reconstructed without unnecessary replication of telemetry information.

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 July 23, 2021.

Copyright Notice

Copyright (c) 2021 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 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.

Table of Contents

1. Introduction	2
2. Requirements for Multicast Traffic Telemetry	3
3. Issues of Existing Techniques	4
4. Proposed Modifications to Existing Techniques	4
4.1. Per-hop postcard using IOAM DEX	5
4.2. Per-section postcard	7
5. Considerations for Different Multicast Protocols	8
5.1. Application in PIM	8
5.2. Application in P2MP	9
5.3. Application in BIER	9
6. Security Considerations	10
7. IANA Considerations	10
8. Contributors	10
9. Acknowledgments	10
10. References	10
10.1. Normative References	10
10.2. Informative References	11
Authors' Addresses	12

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.ietf-ippm-ioam-data], Postcard-based Telemetry

[I-D.song-ippm-postcard-based-telemetry], and Hybrid Two-Step (HTS) [I-D.mirsky-ippm-hybrid-two-step] 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.

When a network contains multicast (p2mp) trees there will be redundant data as data is replicated at branch points. The IP Multicast S,G data is identical from one branch to another on it's way to multiple receivers. When adding iOAM trace data, to multicast packets, we enlarge data packets thus consuming more network bandwidth. Instead of adding iOAM trace data, it could be more efficient to collect the telemetry information using solutions, such as iOAM postcard or HTS, to cut down on the redundant iOAM data. The problem is that a postcard type solution doesn't have a branch identifier.

This draft proposes a set of solutions to this iOAM data redundancy problem. The requirements for multicast traffic telemetry are discussed along with the issues of the existing on-path telemetry techniques. We propose modifications to make these techniques adapt to multicast in order for the original multicast tree to be correctly reconstructed while eliminating redundant data.

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, a few of which are:

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

In order to meet these requirements, we need the ability to directly monitor the multicast traffic and derive data from the multicast packets. The conventional OAM mechanisms, such as multicast ping and trace, may not be sufficient 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) Trace option [I-D.ietf-ippm-ioam-data], IOAM Direct Export (DEX) option [I-D.ioamteam-ippm-ioam-direct-export], and Postcard-based Telemetry with Packet Marking (PBT-M) [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 the IOAM trace option 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, however, 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.

The PBT solutions, including the IOAM DEX option and PBT-M, can be used to eliminate such data redundancy, because each node on the tree only sends a postcard covering local data. However, they 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 the other to node D, and Node B leads to Node C and Node D leads to Node E. From the received postcards, 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

Two solutions are proposed to address the above issues. One is built on PBT and requires augmentation or modification to the instruction header of the IOAM Direct Export Option; the other combines the IOAM trace option and PBT for an optimized solution.

4.1. Per-hop postcard using IOAM DEX

One way to mitigate PBT's multiple tree tracking weakness is to augment it with a branch identifier field. Note that this works for the IOAM DEX option but not for PBT-M because the IOAM DEX option uses an instruction header. To make the branch identifier globally unique, the branch node ID plus an index is used. For example, if Node A has two branches, one to Node B and one to Node 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 for each multicast tree instance 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, along with the other fields such as flow ID and sequence number, is 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 contains the branch identifier. The curly brace contains the telemetry data about a specific node.

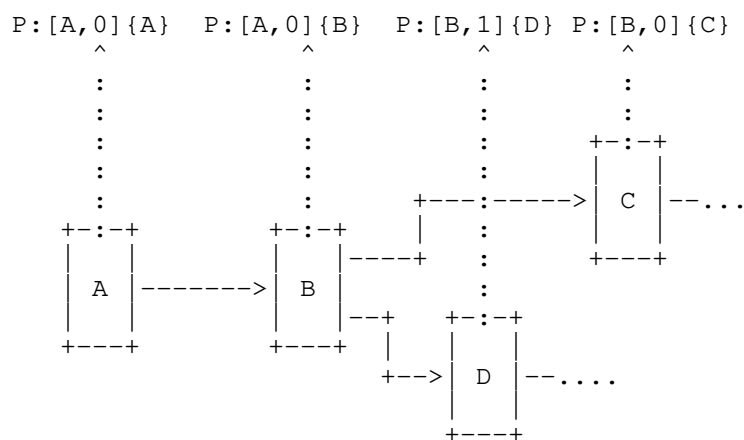


Figure 1: Per-hop Postcard

Each branch fork node need to generate the branch ID for each branch in its multicast tree instance and include it in the IOAM DEX option header so the downstream node can learn it. The branch ID contains two parts: the branch fork node ID and a unique branch index.

Figure 2 shows that the branch ID is carried as an optional field after the flow ID and sequence number optional fields in the IOAM DEX option header. A bit "M" in the Flags field is reserved to indicate

the presence of the branch index field. The "M" flag position will be determined later after the other flags are specified in [I-D.ioamteam-ippm-ioam-direct-export].

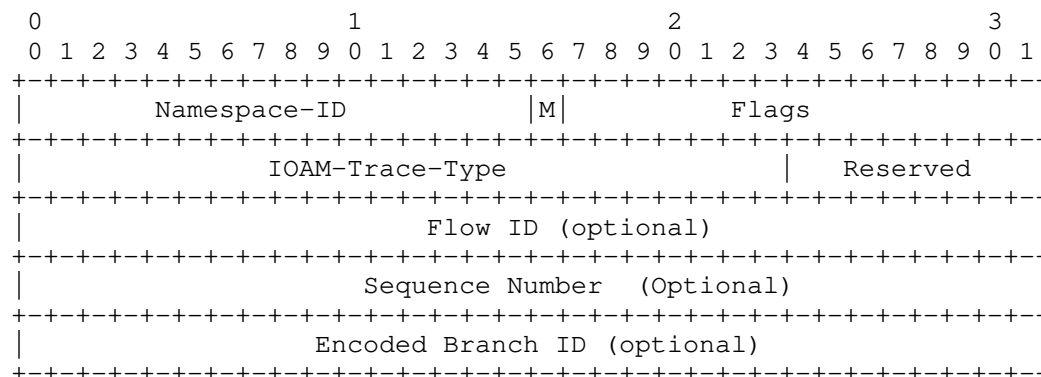


Figure 2: Carry Branch Index in IOAM DEX option header

To avoid introducing a new type of data field to the IOAM DEX option header, we can encode the branch identifier using the existing node ID data field as defined in [I-D.ietf-ippm-ioam-data]. Currently, the node ID field occupies three octets. A simple solution is to shorten the node ID field so a number of bits can be saved to encode the branch index, as shown in Figure 3.

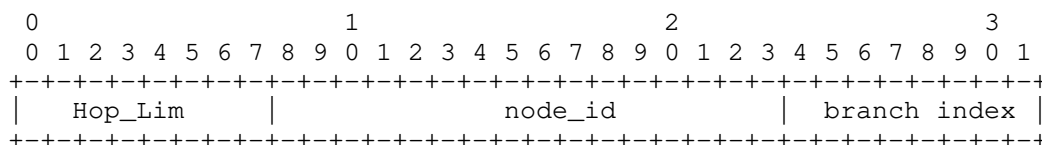


Figure 3: Encode Branch Index with Node ID Method 1

Another encoding method is to use the sum of the node ID and the branch index as the new node ID, as shown in Figure 4. As long as the node IDs are assigned with large enough gap, the telemetry data analyzer can still successfully recover the original node ID and branch index.

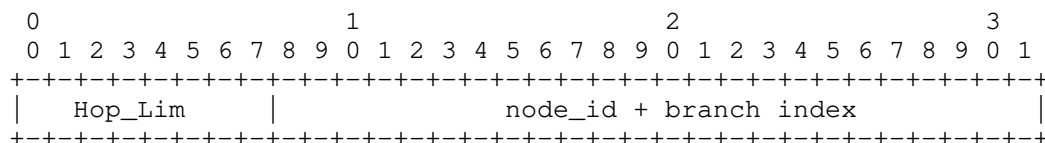


Figure 4: Encode Branch Index with Node ID Method 2

Once a node gets the branch ID information from the upstream, it MUST carry this information in its telemetry data export postcards, so the original multicast tree can be correctly reconstructed based on the postcards.

4.2. Per-section postcard

The second solution is a combination of the IOAM trace mode and PBT. To avoid data redundancy at each branch node, the trace data accumulated, to that point, is exported by a postcard before the packet is replicated. In this case, each branch still needs to maintain some identifier to help correlate the postcards for each tree section. The natural way to accomplish this 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, the local data exported by each branch node can only contain partial data (e.g., ingress interface and ingress timestamp).

Figure 5 shows an example in a segment of a multicast tree. Node B and D are two branch nodes and they will export a postcard covering the trace data for the previous section. The end node of each path will also need to export the data of the last section as a postcard.

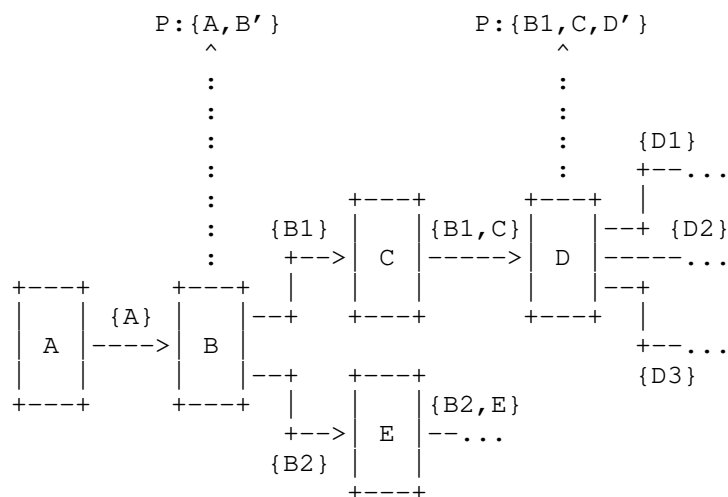


Figure 5: Per-section Postcard

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

The authors would like to thank Frank Brockners, Tianran Zhou for the comments and advice.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4687] Yasukawa, S., Farrel, A., King, D., and T. Nadeau, "Operations and Management (OAM) Requirements for Point-to-Multipoint MPLS Networks", RFC 4687, DOI 10.17487/RFC4687, September 2006, <<https://www.rfc-editor.org/info/rfc4687>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <<https://www.rfc-editor.org/info/rfc7761>>.
- [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>>.
- [RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <<https://www.rfc-editor.org/info/rfc8279>>.

[RFC8487] Asaeda, H., Meyer, K., and W. Lee. Ed., "Mtrace Version 2: Traceroute Facility for IP Multicast", RFC 8487, DOI 10.17487/RFC8487, October 2018, <<https://www.rfc-editor.org/info/rfc8487>>.

10.2. Informative References

- [I-D.herbert-ipv4-udpencap-eh]
Herbert, T., "IPv4 Extension Headers and UDP Encapsulated Extension Headers", draft-herbert-ipv4-udpencap-eh-01 (work in progress), March 2019.
- [I-D.ietf-bier-oam-requirements]
Mirsky, G., Nainar, N., Chen, M., and S. Pallagatti, "Operations, Administration and Maintenance (OAM) Requirements for Bit Index Explicit Replication (BIER) Layer", draft-ietf-bier-oam-requirements-11 (work in progress), November 2020.
- [I-D.ietf-ippm-ioam-data]
Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In-situ OAM", draft-ietf-ippm-ioam-data-11 (work in progress), November 2020.
- [I-D.ioametal-ippm-6man-ioam-ipv6-deployment]
Bhandari, S., Brockners, F., Mizrahi, T., Kfir, A., Gafni, B., Spiegel, M., Krishnan, S., and M. Smith, "Deployment Considerations for In-situ OAM with IPv6 Options", draft-ioametal-ippm-6man-ioam-ipv6-deployment-03 (work in progress), March 2020.
- [I-D.ioamteam-ippm-ioam-direct-export]
Song, H., Gafni, B., Zhou, T., Li, Z., Brockners, F., Bhandari, S., Sivakolundu, R., and T. Mizrahi, "In-situ OAM Direct Exporting", draft-ioamteam-ippm-ioam-direct-export-00 (work in progress), October 2019.
- [I-D.mirsky-ippm-hybrid-two-step]
Mirsky, G., Lingqiang, W., Zhui, G., and H. Song, "Hybrid Two-Step Performance Measurement Method", draft-mirsky-ippm-hybrid-two-step-07 (work in progress), December 2020.
- [I-D.song-ippm-postcard-based-telemetry]
Song, H., Zhou, T., Li, Z., Mirsky, G., Shin, J., and K. Lee, "Postcard-based On-Path Flow Data Telemetry using Packet Marking", draft-song-ippm-postcard-based-telemetry-08 (work in progress), October 2020.

[I-D.song-mpls-extension-header]

Song, H., Li, Z., Zhou, T., and L. Andersson, "MPLS Extension Header", draft-song-mpls-extension-header-02 (work in progress), February 2019.

[I-D.xie-bier-ipv6-encapsulation]

Xie, J., Geng, L., McBride, M., Asati, R., Dhanaraj, S., Zhu, Y., Qin, Z., Shin, M., Mishra, G., and X. Geng, "Encapsulation for BIER in Non-MPLS IPv6 Networks", draft-xie-bier-ipv6-encapsulation-09 (work in progress), January 2021.

Authors' Addresses

Haoyu Song
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

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Gyan Mishra
Verizon Inc.

Email: gyan.s.mishra@verizon.com

PIM Working Group
Internet Draft
Intended status: Standards Track
Expires: January 02, 2020

H. Zhao
Ericsson
X. Liu
Volta
Y. Liu
Huawei
M. Panchanathan
Cisco
M. Sivakumar
Juniper

July 03, 2019

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/lid-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 January 02, 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 (<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.

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:

- o Brackets "[" and "]" enclose list keys.
- o Abbreviations before data node names: "rw" means configuration (read-write), and "ro" means state data (read-only).
- o Symbols after data node names: "?" means an optional node, "!" means a presence container, and "*" denotes a list and leaf-list.
- o Parentheses enclose choice and case nodes, and case nodes are also marked with a colon (":").
- o 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
```

```

augment /rt:routing/rt:control-plane-protocols/rt:control-plane-protocol:
  +--rw mld-proxy {feature-mld-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:ipv6-address
          +--ro up-time?           uint32
          +--ro filter-mode?       enumeration
          +--ro source* [source-address]
            +--ro source-address    inet:ipv6-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;
  namespace "urn:ietf:params:xml:ns:yang:ietf-igmp-mld-proxy";
  // 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";

```

contact

"WG Web: <<http://tools.ietf.org/wg/pim/>>
WG List: <<mailto:pim@ietf.org>>

Editors: Hongji Zhao
<<mailto:hongji.zhao@ericsson.com>>

Xufeng Liu
<<mailto:xufeng.liu.ietf@gmail.com>>

Yisong Liu
<<mailto:liuyisong@huawei.com>>

Mani Panchanathan
<<mailto:mapancha@cisco.com>>

Mahesh Sivakumar
<<mailto:sivakumar.mahesh@gmail.com>>

";

description

"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";
  }
}
```

```
} // per-interface-config-attributes

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.";
  }
} // state-group-attributes

/* 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";
  }

  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
      "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
```

```
        "Filter mode for a multicast group,  
        may be either include or exclude.";  
    }  
    }  
    } // list source  
    } // list group  
    } // interface  
    } // interfaces  
}  
}  
  
/*  RPCs  */  
  
}  
<CODE ENDS>
```

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]:

URI: urn:ietf:params:xml:ns:yang:ietf-igmp-mld-proxy

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

namespace: urn:ietf:params:xml:ns:yang:ietf-igmp-mld-proxy

prefix: imp

reference: RFC XXXX

6. Normative References

- [RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2236, November 1997.
- [RFC2710] Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, October 1999.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, August 2006.
- [RFC4605] B. Fenner, H. He, B. Haberman and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD) - Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.
- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.
- [RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, October 2010.
- [RFC6991] Schoenwaelder, J., Ed., "Common YANG Data Types", RFC 6991, July 2013.
- [RFC8342] M. Bjorklund and J. Schoenwaelder, "Network Management Datastore Architecture (NMDA)", RFC 8342, March 2018.
- [RFC8343] M. Bjorklund, "A YANG Data Model for Interface Management", RFC 8343, March 2018.
- [draft-ietf-pim-igmp-mld-yang-06] X. Liu, F. Guo, M. Sivakumar, P. McAllister, A. Peter, "A YANG data model for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD)", draft-ietf-pim-igmp-mld-yang-06, Oct 20, 2017.
- [draft-dsdt-nmda-guidelines-01] M. Bjorklund, J. Schoenwaelder, P. Shafer, K. Watsen, R. Wilton, "Guidelines for YANG Module Authors (NMDA)", draft-dsdt-nmda-guidelines-01, May 2017

[draft-ietf-netmod-revised-datastores-03] M. Bjorklund, J.
Schoenwaelder, P. Shafer, K. Watsen, R. Wilton, "Network
Management Datastore Architecture", draft-ietf-netmod-
revised-datastores-03, July 3, 2017

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 Beiqing 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

