

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: April 22, 2011

D. Zhou, Ed.
Hangzhou H3C Tech. Co., Ltd.
H. Deng
China Mobile Research Institute
Y. Shi
Hangzhou H3C Tech. Co., Ltd.
H. Liu
Huawei Technologies Co., Ltd.
I. Bhattacharya
Cisco Systems
October 19, 2010

Unnecessary Multicast Flooding Problem Statement
draft-dizhou-pim-umf-problem-statement-01

Abstract

This document describes the unnecessary multicast stream flooding problem in the link layer switches between multicast source and PIM First Hop Router (FHR). The IGMP-Snooping Switch will forward multicast streams to router ports, and the PIM FHR must receive all multicast streams even if there is no request from receiver. This often leads to waste of switches' cache and link bandwidth when the multicast streams are not actually required. This document details the problem and defines design goals for a generic mechanism to restrain the unnecessary multicast stream flooding.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 22, 2011.

Copyright Notice

Copyright (c) 2010 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.

1. Introduction

The Protocol Independent Multicast (PIM) is now the most popular multicast routing protocol in the world. The router on the edge of PIM routing domain is called PIM First Hop Router (FHR). PIM has four modes: Sparse Mode (SM), Dense Mode (DM), Bidirectional PIM (Bidir-PIM), Source-Specific Multicast (SSM). DM and Bidir-PIM suppose that the receivers are always existing. SM and SSM are designed for multicast streams to be transferred on demand.

The IGMP-Snooping, specified in RFC 4541 [RFC4541], is a link layer multicast streams forwarding control mechanism. It forwards all multicast streams to the router ports and selected multicast stream to membership ports. It also forwards IGMP Membership report messages to router ports and IGMP Query messages to all ports except the incoming port.

The PIM-Snooping is another link layer multicast streams forwarding control mechanism. It forwards the selected multicast streams to the requested router ports. But it can not run on the path between multicast source and PIM FHR because there is no PIM Join and PIM Prune messages.

In many typical deployment scenarios, some link layer switches are existing between multicast sources and PIM FHR. The receivers may be only exist between the source and PIM FHR, maybe exist in the network behind the PIM FHR, maybe even not exist temporarily. But if only the PIM FHR exists, the multicast streams are always transferred through these switches to PIM FHR, even if no receivers exist.

These unnecessary multicast streams will lead to waste of the switches' cache and link bandwidth. And the cache and link bandwidth are essential for application streams to transfer with less packet loss, latency, and jitter.

There are some attempts made to solve the problem of unnecessary multicast streams flooding on switches between the sources and PIM FHR in various ways. However, those solutions are either scenario-limited or deployment-limited.

This document provides a detailed description of protocol design goals for efficient PIM and PIM-Snooping based mechanism to solve this problem.

2. Terminology

In this document, several words are used to signify the requirements of the specification. These words are often capitalized. 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]

With respect to PIM, this document follows the terminology that has been defined in RFC 4601 [RFC4601].

3. Problem Statement

Under the existing model, the PIM FHR sends out PIM hello messages, as well as IGMP query messages if it is an IGMP querier in the network segment. The IGMP-Snooping switches between the sources and PIM FHR receive multicast streams from the sources and forward them to PIM FHR, even there is no request from receivers. If there are many sources, the multicast streams will flood to all PIM FHRs.

This will bring about some serious problems.

Firstly, the unnecessary multicast streams will seriously consume link bandwidth .

If there are many sources and PIM FHRs, the bandwidth between the source and PIM FHR will be seriously wasted. This problem will be more serious especially in such a type of application:

- o The sources are much more than receivers.
- o Each source may be requested simultaneously by some receivers.
- o Most sources are NOT requested at most time.
- o Receivers may be only exist between the source and PIM FHR, maybe exist in the network behind the PIM FHR, maybe even not exist temporarily.

In the application mentioned above, it is not acceptable to afford plenty of bandwidth to forward all multicast streams from the sources.

Secondly, the unnecessary multicast streams will consume outgress cache of switches.

In any network deployment, the switch between the sources and PIM FHR forwarding unnecessary multicast streams will also consume the outgress cache of switch including out-port specified cache and the cache shared by all ports. The cache is the key resource of switch to reduce streams' packet loss ratio, latency, and jitter.

Finally, the unnecessary multicast streams forwarding will increase the power consumption.

In summary, it is desirable to afford a mechanism to prohibit the switches between the source and PIM FHR from forwarding unnecessary multicast stream when it is not requested, and drive the switches to forward multicast stream in time when it is required.

4. Design Goals

The following are the goals and constraints in designing the mechanism for switch to restrain unnecessary multicast streams flooding:

- o Switch SHALL forward the requested streams and SHALL NOT forward unrequired streams.
- o Streams SHALL just be terminated at the exact switch.
- o If a receiver appears, it MUST receive multicast streams in time.
- o Deployment SHALL be flexible. The number and topology of switches between source and PIM FHR SHALL NOT be limited. The ip address deployment of multicast sources and receivers SHALL NOT be limited either. Sources and receivers may be in the same ip address segment, for example.
- o The CPUs of switches SHALL receive no multicast stream data, but only protocol messages.

5. Use Cases and Related Work

In order to further clarify the items listed in scope of the proposed work, this section provides some background on related work and the use cases envisioned for the proposed work.

5.1. Source sending stream on demand

By adding a central controlling server, the multicast sources may be controlled to send streams on demand.

Note that once a receiver sends a request, the multicast stream will flow down toward switch's router ports, even if there is no other receivers behind the router ports.

5.2. Host simulation of PIM FHR

PIM FHR may be prohibited to send PIM hello messages and IGMP Query messages toward multicast sources. Instead, it can simulate host to send IGMP Membership Report and Leave messages if it receives PIM Join and Prune messages. So the switches between the sources and PIM FHRs would have no router ports.

But for PIM SM, the PIM FHR does not know at which port to send out IGMP messages, unless configured some information at the requested ports by network manager.

On the other hand, the switch will not forward IGMP Membership Report and Leave messages towards sources. It will only forward IGMP Membership Report and Leave messages to router ports.

5.3. IGMP Querier simulation of first-hop switch

For the second problem of PIM FHR host simulation, the switch directly connected to source can simulate IGMP Querier.

But when there are two or more switches simulating IGMP Queriers, the phenomenon of unnecessary multicast streams flooding still exists.

5.4. Replacing link layer switches with Routers

Replacing link layer switches directly connected to sources with Routers is not a perfect solution either. It will limit the flexibility of networking, and will further lead to waste of ip address and many ip address segments seriously if there are many sources. It will also bring about many ip address segments and then complicate network management.

6. some potential solutions

6.1. solution based on PIM and PIM-Snooping

The key points of it are as follows:

- o When the PIM FHR receives a multicast stream, it creates an entry of (S,G) if the entry did not exist. And it judges whether the (S,G) entry has out interfaces. If the (S,G) has no out interface, the PIM router sends out a unicast PIM prune message towards the multicast source. The upstream neighbor address in the message is the source address.
- o The switch between multicast sources and PIM FHR runs PIM snooping and IGMP snooping. When it intercepts the unicast PIM prune message by ip protocol field identification and finds out that the upstream neighbor address of the message is not in its PIM neighbor lists, it creates a (S,G) entry with a pruned port and an upstream port if not created before. The upstream port is found by looking up the unicast mac table. That (S,G) entry is punched with a specific sign which means that entry is different from traditional PIM-Snooping entry. The pruned port SHALL NOT forward multicast stream and has a lifetime which is 1/3 of that of PIM FHR's (S,G) entry, then converted to be a downstream port, so that the multicast stream will arrive at PIM FHR to refresh the (S,G) entry.
- o Looking up IGMP-snooping entry and PIM-snooping entry, if the switch find there is no need to forward the multicast stream, it SHALL forward the unicast PIM prune message towards multicast source.
- o When the switch receives an IGMP membership report, it shall forward the message through its router ports and upstream port.
- o When PIM FHR creates an out interface for a (S,G) entry that had no out interface before, it shall send unicast PIM join message towards multicast source. The upstream neighbor address of the message is the source address.
- o When the switch receives the unicast PIM join message and finds out that the upstream neighbor address of the message is not in its PIM neighbor lists, it will convert the pruned port to be downstream port. When the (S,G) entry with specific sign has no pruned ports, it should be deleted in order to save the entry space.

- o By the information from IGMP-snooping entry and PIM-snooping entry, the switch can decide whether it shall forward the unicast PIM join message towards multicast source.
- o The role of membership port is prior than that of pruned port, and the role of pruned port is prior than that of router port or downstream port.
- o If two or more switches or PIM FHRs are connected by one port directly, or through HUB or normal switch, some query mechanism shall be implemented.

6.2. solution based on IGMP and IGMP-Snooping

The key points of it are as follows:

- o When the PIM FHR receives a multicast stream, it creates an entry of (S,G) if the entry did not exist. And it judges whether the (S,G) entry has out interfaces. If the (S,G) has no out interface, the PIM router sends out an unicast IGMP prune message towards multicast source.
- o The switch between multicast sources and PIM FHR runs IGMP snooping. When it intercepts the unicast IGMP prune message by ip protocol field identification, it creates a IGMP-Snooping entry with a pruned port and an source port. The source port is found by looking up the unicast mac table. The pruned port has a lifetime which is 1/3 of the lifetime of PIM FHR's (S,G) entry, so that the multicast stream will arrive at PIM FHR before its (S,G) entry dies out.
- o By the information from IGMP-snooping entry, the switch can decide whether it shall forward the unicast IGMP prune message towards multicast source.
- o When the switch receives an IGMP membership report, it shall forward the message through its router ports and source port.
- o When PIM FHR creates an out interface for a (S,G) entry that had no out interface before, it shall send unicast IGMP graft message towards multicast source.
- o When the switch receives the unicast IGMP graft message, it will change the pruned port to be router port. When the IGMP-Snooping entry has only router ports and source ports, it should be deleted in order to save the entry space.

- o By the information from IGMP-snooping entry, the switch can decide whether it shall forward the unicast IGMP graft message towards multicast source.
- o The role of membership port is prior than that of pruned port, and the role of pruned port is prior than that of router port.
- o If two or more switches or PIM FHRs are connected by one port directly, or through HUB or normal switch, some query mechanism shall be implemented.

The first solution is more simple than the second one because the PIM-snooping has afforded some essential information and there is no need to add some new messages. In the first solution the switches must run PIM-snooping besides IGMP-snooping.

Any advice is welcome.

7. Security Considerations

8. Contributors

9. Acknowledgements

10. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, May 2006.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

Authors' Addresses

Di Zhou (editor)
Hangzhou H3C Tech. Co., Ltd.
310 Liuhe Road
Hangzhou, Zhejiang
China(310053)

Phone: +86-571-86761327
Email: zhouidi@h3c.com

Hui Deng
China Mobile Research Institute
Unit2,28 Xuanwumenxi Ave,Xuanwu District
Beijing, Beijing
China(100053)

Phone: +86-010-15801696688-3314
Email: denghui@chinamobile.com

Yang Shi
Hangzhou H3C Tech. Co., Ltd.
Beijing R&D Center of H3C, Digital Technology Plaza,
NO.9 Shangdi 9th Street,Haidian District,
Beijing
China(100085)

Phone: +86 010 82775276
Email: young@h3c.com

Hui Liu
Huawei Technologies Co., Ltd.
Huawei Bld., No.3 Xinxu Rd.
Shang-Di Information Industry Base, Hai-Dian District,
Beijing
China(100085)

Email: Liuhui47967@huawei.com

Indranil Bhattacharya
Cisco Systems
India(560037)

Email: myselfindranil@gmail.com

PIM Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 22, 2011

B. Joshi
Infosys Technologies Ltd.
A. Kessler
Cisco Systems, Inc.
D. McWalter
July 21, 2010

PIM Group-to-RP Mapping
draft-ietf-pim-group-rp-mapping-05.txt

Abstract

Each PIM-SM router in a PIM Domain which supports ASM maintains Group-to-RP mappings which are used to identify a RP for a specific multicast group. PIM-SM has defined an algorithm to choose a RP from the Group-to-RP mappings learned using various mechanisms. This algorithm does not consider the PIM mode and the mechanism through which a Group-to-RP mapping was learned.

This document defines a standard algorithm to deterministically choose between several group-to-rp mappings for a specific group. This document first explains the requirements to extend the Group-to-RP mapping algorithm and then proposes the new algorithm.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 22, 2011.

Copyright Notice

Copyright (c) 2010 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
2. Terminology	4
3. Existing algorithm	5
4. Assumptions	6
5. Common use cases	7
6. Proposed algorithm	8
7. Deprecation of MIB Objects	10
8. Clarification for MIB Objects	11
9. Use of dynamic group-to-rp mapping protocols	12
10. Consideration for Bidirectional-PIM and BSR hash	13
11. Filtering Group-to-RP mappings at domain boundaries	14
12. Security Consideration	15
13. IANA Consideration	16
14. Acknowledgements	17
15. Normative References	18
Authors' Addresses	19

1. Introduction

Multiple mechanisms exist today to create and distribute Group-to-RP mappings. Each PIM-SM router may learn Group-to-RP mappings through various mechanisms.

It is critical that each router select the same 'RP' for a specific multicast group address. This is even true in the case of Anycast RP for redundancy. This RP address may correspond to a different physical router but it is one logical RP address and must be consistent across the PIM domain. This is usually achieved by using the same algorithm to select the RP in all the PIM routers in a domain.

PIM-SM [RFC4601] has defined an algorithm to select a 'RP' for a given multicast group address but it is not flexible enough for an administrator to apply various policies. Please refer to section 3 for more details.

PIM-STD-MIB [RFC5060] has defined an algorithm that allows administrators to override Group-to-RP mappings with static configuration. But this algorithm is not completely deterministic, because it includes an implementation-specific 'precedence' value.

Embedded-RP as defined in section-7.1 of Embedded-RP address in IPv6 Multicast address [RFC3956], mentions that to avoid loops and inconsistencies, for addresses in the range FF70::/12, the Embedded-RP mapping must be considered the longest possible match and higher priority than any other mechanism.

2. Terminology

In this document, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in RFC 2119 [RFC2119]. This document also uses following terms:

- o PIM Mode

PIM Mode is the mode of operation a particular multicast group is used for. Wherever this term is used in this document, it refers to either Sparse Mode or BIDIR Mode.

- o Dynamic group-to-RP mapping mechanisms

The term Dynamic group-to-RP mapping mechanisms in this document refers to BSR and Auto-RP.

- o Dynamic mappings or Dynamically learned mappings

The terms Dynamic mappings or Dynamically learned mappings refer to group-to-RP mappings that have been learned by BSR or Auto-RP. Group-to-RP mappings that have been learned by embedded RP are referred to as Embedded Group-to-RP mappings.

- o Filtering

Filtering is selective discarding of dynamic Group-to-RP mapping information, based on the group address, the type of Group-to-RP mapping message and the interface on which the mapping message was received.

- o Multicast Domain and Boundaries

The term multicast domain used in this document refers to a network topology that has a consistent set of Group-to-RP Mappings. The interface between two or more multicast domains is a multicast domain boundary. The multicast boundaries are usually enforced by filtering the dynamic mapping messages and/or configuring different static RP mappings.

3. Existing algorithm

Existing algorithm defined in PIM-SM (Section 4.7.1 in [RFC4601]) does not consider following constraints:

- o It does not consider the origin of a Group-to-RP mapping and therefore will treat all of them equally.
- o It does not provide the flexibility to give higher priority to a specific PIM mode. For example, an entry learned for PIM-BIDIR mode is treated with same priority as an entry learned for PIM-SM.

4. Assumptions

We have made following assumptions in defining this algorithm:

- o Embedded Group-to-RP mappings are special and always have the highest priority. They cannot be overridden either by static configuration or by dynamic Group-to-RP mappings.
- o Dynamic mappings will override a static RP config if they have overlapping ranges. However, it is possible to override dynamic Group-to-RP mappings with static configurations, either by filtering, or by configuring longer static group addresses that override dynamic mappings when longest prefix matching is applied.
- o A Group-to-RP mapping can be learned from various mechanisms. We assume that following list is in the decreasing preferences of these mechanism:
 - * Embedded Group-to-RP mappings
 - * Dynamically learned mappings
 - * Static configuration.
 - * Other mapping method
- o A Group-to-RP mapping learned for PIM-BIDIR mode is preferred to an entry learned for PIM-SM mode.
- o Dynamic group-to-RP mapping mechanisms are filtered at domain boundaries or for policy enforcement inside a domain.

5. Common use cases

- o Default static Group-to-RP mappings with dynamically learned entries

Many network operators will have a dedicated infrastructure for the standard multicast group range (224/4) and so might be using statically configured Group-to-RP mappings for this range. In this case, to support some specific applications, they might like to learn Group-to-RP mappings dynamically using either BSR or Auto-RP mechanism. In this case to select Group-to-RP mappings for these specific applications, a longer prefix match should be given preference over statically configured Group-to-RP mappings. For example 239.100.0.0/16, an administratively scoped multicast address range, could be learned for a corporate communications application. Network operators may change the Group-to-RP mappings for these applications more often and would need to be learned dynamically.

- o Migration situations

Network operators occasionally go through a migration due to an acquisition or a change in their network design. In order to facilitate this migration there is a need to have a deterministic behaviour of Group-to-RP mapping selection for entries learned using BSR and Auto-RP mechanism. This will help in avoiding any unforeseen interoperability issues between different vendor's network elements.

- o Use by management systems

A network management station can determine the RP for a specific group in a specific router by running this algorithm on the Group-to-RP mapping table fetched using SNMP MIB objects.

- o More use cases

By no means, the above list is complete. Please drop a mail to 'authors' if you see any other use case for this.

6. Proposed algorithm

The following algorithm addresses the above mentioned shortcomings in the existing mechanism:

1. If the Multicast Group Address being looked up contains an embedded RP, RP address extracted from the Group address is selected as Group-to-RP mapping.
2. If the Multicast Group Address being looked up is in the SSM range or is configured for Dense mode, no Group-to-RP mapping is selected, and this algorithm terminates. Alternatively, a RP with address type 'unknown' can be selected. Please look at section #8 for more details on this.
3. From the set of all Group-to-RP mapping entries, the subset whose group prefix contains the multicast group that is being looked up, are selected.
4. If there are no entries available, then the Group-to-RP mapping is undefined.
5. A longest prefix match is performed on the subset of Group-to-RP Mappings.
 - * If there is only one entry available then that is selected as Group-to-RP mapping.
 - * If there are multiple entries available, we continue with the algorithm with this smaller set of Group-to-RP Mappings.
6. From the remaining set of Group-to-RP Mappings we select the subset of entries based on the preference for the PIM modes which they are assigned. A Group-to-RP mapping entry with PIM Mode 'BIDIR' will be preferred to an entry with PIM Mode 'PIM-SM'.
 - * If there is only one entry available then that is selected as Group-to-RP mapping.
 - * If there are multiple entries available, we continue with the algorithm with this smaller set of Group-to-RP Mappings.
7. From the remaining set of Group-to-RP Mappings we select the subset of the entries based on the origin. Group-to-RP mappings learned dynamically are preferred over static mappings. If the remaining dynamic Group-to-RP mappings are from BSR and Auto-RP then the mappings from BSR SHOULD be preferred.

- * If there is only one entry available then that is selected as Group-to-RP mapping.
 - * If there are multiple entries available, we continue with the algorithm with this smaller set of Group-to-RP Mappings.
8. If the remaining Group-to-RP mappings were learned through BSR then the RP will be selected by comparing the RP Priority in the Candidate-RP-Advertisement messages. The RP mapping with the lowest value indicates the highest priority [RFC5059].
- * If more than one RP has the same highest priority value we continue with the algorithm with those Group-to-RP mappings.
 - * If the remaining Group-to-RP mappings were NOT learned from BSR we continue the algorithm with the next step.
9. If the remaining Group-to-RP mappings were learned through BSR and the PIM Mode of the Group is 'PIM-SM' then the hash function will be used to choose the RP. The RP with the highest resulting hash value will be selected.
- * If more than one RP has the same highest hash value we continue with the algorithm with those Group-to-RP mappings.
 - * If the remaining Group-to-RP mappings were NOT learned from BSR we continue the algorithm with the next step.
10. From the remaining set of Group-to-RP Mappings we will select the RP with the highest IP address. This will serve as a final tiebreaker.

7. Deprecation of MIB Objects

Group-to-RP mapping algorithm defined in PIM-STD-MIB [RFC5060] does not specify the usage of 'pimGroupMappingPrecedence' and 'pimStaticRPPPrecedence' objects in 'pimGroupMappingTable' table clearly. With the newly proposed algorithm in this document, these MIB objects would not be required. So we propose to deprecate these MIB objects from PIM-STD-MIB. Also the newly proposed algorithm in this document MUST be preferred over Group-to-RP mapping algorithm defined in either PIM-SM[RFC4601] or in PIM-STD-MIB[RFC5060].

8. Clarification for MIB Objects

When an Group-to-RP mapping entry is created in the `pimGroupMappingTable` in the PIM-STD MIB[RFC5060], it would be acceptable to have an entry with an RP with address type 'unknown' and a `PimMode` of Dense Mode or SSM. These entries would represent group ranges for Dense mode or SSM.

Also all the entries which are already included in the SSM Range table in the IP Mcast MIB would be copied over to `pimGroupMappingTable`. They would have a type of `configSSM` and an RP with address type 'unknown' as described above.

The advantage of keeping all the ranges in the table would be that this table will contain all the known multicast group ranges.

9. Use of dynamic group-to-rp mapping protocols

In practice, it is not usually necessary to run several dynamic Group-to-RP mapping mechanisms in one administrative domain. Specifically, interoperation of BSR and Auto-RP is OPTIONAL and not recommended by this document.

However, if a router does receive two overlapping sets of Group-to-RP mappings, for example from Auto-RP and BSR, then some algorithm is needed to deterministically resolve the situation. The algorithm in this document **MUST** be used. This can be important at domain border routers, and is likely to improve stability under misconfiguration and when configuration is changing.

An implementation of PIM that supports only one mechanism for learning Group-to-RP mappings **SHOULD** also use this algorithm. The algorithm has been chosen so that existing standard implementations are already compliant.

10. Consideration for Bidirectional-PIM and BSR hash

Bidir-PIM [RFC5015] is designed to avoid any data driven events. This is especially true in the case of a source only branch. The RP mapping is determined based on a group mask when the mapping is received through a dynamic mapping protocol or statically configured.

Therefore the hash in BSR is ignored for PIM-Bidir RP mappings based on the algorithm defined in this document. It is RECOMMENDED that network operators configure only one PIM-Bidir RP for each RP Priority.

11. Filtering Group-to-RP mappings at domain boundaries

An implementation of PIM SHOULD support configuration to block specific dynamic mechanism for a valid group prefix range. For example, it should be possible to allow an administratively scoped address range, such as 239/8 range, for Auto-RP protocol but block the BSR advertisement for the same range. Similarly it should be possible to filter out all Group-to-RP mappings learned from BSR or Auto-RP protocol.

12. Security Consideration

This document does not suggest any protocol specific functionality so there is no security related consideration.

13. IANA Consideration

This draft does not create any namespace for IANA to manage.

14. Acknowledgements

This draft is created based on the discussion occurred during the PIM-STD-MIB [RFC5060] work. Many thanks to Stig Vennas, Yiqun Cai and Toerless Eckert for providing useful comments.

15. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC5060] Sivaramu, R., Lingard, J., McWalter, D., Joshi, B., and A. Kessler, "Protocol Independent Multicast MIB", RFC 5060, January 2008.
- [RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, November 2004.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.
- [RFC5059] Bhaskar, N., Gall, A., Lingard, J., and S. Venaas, "Bootstrap Router (BSR) Mechanism for Protocol Independent Multicast (PIM)", RFC 5059, January 2008.

Authors' Addresses

Bharat Joshi
Infosys Technologies Ltd.
44 Electronics City, Hosur Road
Bangalore 560 100
India

Email: bharat_joshi@infosys.com
URI: <http://www.infosys.com/>

Andy Kessler
Cisco Systems, Inc.
425 E. Tasman Drive
San Jose, CA 95134
USA

Email: kessler@cisco.com
URI: <http://www.cisco.com/>

David McWalter

Email: david@mcwalter.eu

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: April 28, 2011

D. Farinacci
IJ. Wijnands
S. Venaas
cisco Systems
M. Napierala
AT&T Labs
October 25, 2010

A Reliable Transport Mechanism for PIM
draft-ietf-pim-port-04.txt

Abstract

This draft describes how a reliable transport mechanism can be used by the PIM protocol to optimize CPU and bandwidth resource utilization by eliminating periodic Join/Prune message transmission. This draft proposes a modular extension to PIM to use either the TCP or SCTP transport protocol.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 28, 2011.

Copyright Notice

Copyright (c) 2010 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. Requirements Notation	5
1.2. Definitions	5
2. Protocol Overview	6
3. New PIM Hello Options	8
3.1. PIM over the TCP Transport Protocol	8
3.2. PIM over the SCTP Transport Protocol	9
4. Establishing Transport Connections	11
4.1. TCP Connection Maintenance	12
4.2. Moving from PORT to Datagram Mode	13
4.3. On-demand versus Pre-configured Connections	14
4.4. Possible Hello Suppression Considerations	14
4.5. Avoiding a Pair of Connections between Neighbors	15
5. Common Header Definition	16
6. Explicit Tracking	20
7. Multiple Instances and Address-Family Support	21
8. Miscellany	22
9. Security Considerations	23
10. IANA Considerations	24
11. Contributors	25
12. Acknowledgments	26
13. References	27
13.1. Normative References	27
13.2. Informative References	27
Authors' Addresses	28

1. Introduction

The goals of this specification are:

- o To create a simple incremental mechanism to provide reliable PIM message delivery in PIM version 2 for use with PIM Sparse-Mode [RFC4601] (including Source-Specific Multicast) and Bidirectional PIM [RFC5015].
- o The reliable transport mechanism will be used for Join-Prune message transmission only.
- o When a router supports this specification, it need not use the reliable transport mechanism with every neighbor. That is, negotiation on a per neighbor basis will occur.

The explicit non-goals of this specification are:

- o Changes to the PIM message formats as defined in [RFC4601].
- o Provide support for automatic switching between Datagram mode and Transport mode. Two routers that are PIM neighbors on a link will always use Transport mode if and only if both have Transport mode enabled.

This document will specify how periodic Join/Prune message transmission can be eliminated by using TCP [RFC0793] or SCTP [RFC4960] as the reliable transport mechanism for Join/Prune messages.

This specification enables greater scalability in terms of control traffic overhead. However, for routers connected to multi-access links that comes at the price of increased control plane state overhead and the control plane overhead required to maintain this state.

In many existing and emerging networks, particularly wireless and mobile satellite systems, link degradation due to weather, interference, and other impairments can result in temporary spikes in the packet loss. In these environments, periodic PIM joining can cause join latency when messages are lost causing a retransmission only 60 seconds later. By applying a reliable transport, a lost join is retransmitted rapidly. Furthermore, when the last user leaves a multicast group, any lost prune is similarly repaired and the multicast stream is quickly removed from the wireless/satellite link. Without a reliable transport, the multicast transmission could otherwise continue until it timed out, roughly 3 minutes later. As network resources are at a premium in many of these environments,

rapid termination of the multicast stream is critical to maintaining efficient use of bandwidth.

1.1. Requirements Notation

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

1.2. Definitions

PORT: Stands for PIM Over Reliable Transport. Which is the short form for describing the mechanism in this specification where PIM can use the TCP or SCTP transport protocol.

Periodic Join/Prune message: A Join/Prune message sent periodically to refresh state.

Incremental Join/Prune message: A Join/Prune message sent as a result of state creation or deletion events. Also known as a triggered message.

Native Join/Prune message: A Join/Prune message which is carried with an IP protocol type of PIM.

PORT Join/Prune message: A Join/Prune message using TCP or SCTP for transport.

Datagram Mode: The current procedures PIM uses by encapsulating Join/Prune messages in IP packets sent either triggered or periodically.

PORT Mode: Procedures used by PIM defined in this specification for sending Join/Prune messages over the TCP or SCTP transport layer.

2. Protocol Overview

PIM Over Reliable Transport (PORT) is a simple extension to PIMv2 for refresh reduction of PIM Join/Prune messages. It involves sending incremental rather than periodic Join/Prune messages over a TCP/SCTP connection between PIM neighbors.

PORT only applies to PIM Sparse-Mode [RFC4601] and Bidirectional PIM [RFC5015] Join/Prune messages.

This document does not restrict PORT to any specific link types. However, the use of PORT on e.g. multi-access LANs with many PIM neighbors should be carefully evaluated. This due to the fact that there may be a full mesh of PORT connections, and that explicit tracking of all PIM PORT routers is required.

PORT can be incrementally used on a link between PORT capable neighbors. Routers which are not PORT capable can continue to use PIM in Datagram Mode. PORT capability is detected using new PORT Capable PIM Hello Options.

Once PORT is enabled on an interface and a PIM neighbor also announces that it is PORT enabled, only PORT Join/Prune messages will be used. That is, only PORT Join/Prune messages are accepted from, and sent to, that particular neighbor. Native Join/Prune messages may still be used for other neighbors.

PORT Join/Prune messages are sent using a TCP/SCTP connection. When two PIM neighbors are PORT enabled, both for TCP or both for SCTP, they will immediately, or on-demand, establish a connection. If the connection goes down, they will again immediately, or on-demand, try to reestablish the connection. No Join/Prune messages (neither Native nor PORT) are sent while there is no connection.

When PORT is used, only incremental Join/Prune messages are sent from downstream routers to upstream routers. As such, downstream routers do not generate periodic Join/Prune messages for state for which the RPF neighbor is PORT-capable.

For Joins and Prunes, which are received over a TCP/SCTP connection, the upstream router does not start or maintain timers on the outgoing interface entry. Instead, it keeps track of which downstream routers have expressed interest. An interface is deleted from the outgoing interface list only when all downstream routers on the interface, no longer wish to receive traffic.

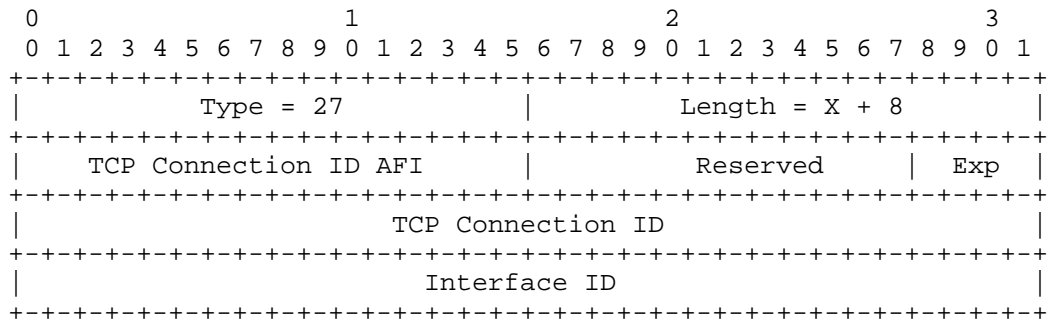
There is no change proposed for the PIM Join/Prune packet format. However, for Join/Prune messages sent over TCP/SCTP connections, no

IP Header is included. The message begins with the PIM common header, followed by the Join/Prune message. See section Section 5 for details on the common header.

3. New PIM Hello Options

3.1. PIM over the TCP Transport Protocol

Option Type: PIM-over-TCP Capable



Allocated Hello Type values can be found in [HELLO-OPT].

When a router is configured to use PIM over TCP on a given interface, it MUST include the PIM-over-TCP Capable hello option in its Hello messages for that interface. If a router is explicitly disabled from using PIM over TCP it MUST NOT include the PIM-over-TCP Capable hello option in its Hello messages. When the router cannot setup a TCP connection, it will refrain from including this option.

Implementations may provide a configuration option to enable or disable PORT functionality. We recommend that this capability be disabled by default.

Length: In bytes for the value part of the Type/Length/Value encoding. Where X is 4 bytes if AFI of value 1 (IPv4) is used and 16 bytes when AFI of value 2 (IPv6) is used [AFI].

TCP Connection ID AFI: The AFI value to describe the address-family of the address of the TCP Connection ID field. When this field is 0, a mechanism outside the scope of this spec is used to obtain the addresses used to establish the TCP connection.

Reserved: Set to zero on transmission and ignored on receipt.

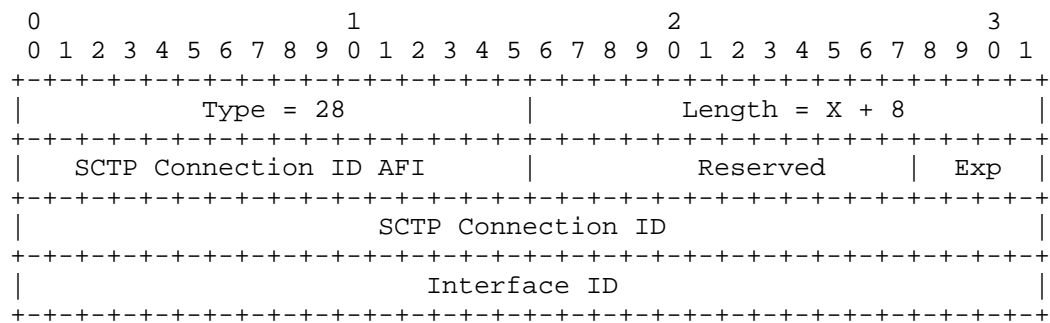
Exp: For experimental use [RFC3692].

TCP Connection ID: An IPv4 or IPv6 address used to establish the TCP connection. This field is omitted (length 0) for the Connection ID AFI 0.

Interface ID: An Interface ID is used to associate the connection a Join/Prune message is received over with an interface which is added or removed from an oif-list. When unnumbered interfaces are used or when a single Transport connection is used for sending and receiving Join/Prune messages over multiple interfaces, the Interface ID is used convey the interface from Join/Prune message sender to Join/Prune message receiver. When a PIM router sets a locally generated value for the Interface ID in the Hello TLV, it must send the same Interface ID value in all Join/Prune messages it is sending to the PIM neighbor.

3.2. PIM over the SCTP Transport Protocol

Option Type: PIM-over-SCTP Capable



Allocated Hello Type values can be found in [HELLO-OPT].

When a router is configured to use PIM over SCTP on a given interface, it MUST include the PIM-over-SCTP Capable hello option in its Hello messages for that interface. If a router is explicitly disabled from using PIM over SCTP it MUST NOT include the PIM-over-SCTP Capable hello option in its Hello messages. When the router cannot setup a SCTP connection, it will refrain from including this option.

Implementations may provide a configuration option to enable or disable PORT functionality. We recommend that this capability be disabled by default.

Length: In bytes for the value part of the Type/Length/Value encoding. Where X is 4 bytes if AFI of value 1 (IPv4) is used and 16 bytes when AFI of value 2 (IPv6) is used [AFI].

SCTP Connection ID AFI: The AFI value to describe the address-family of the address of the SCTP Connection ID field. When this field is 0, a mechanism outside the scope of this spec is used to obtain the addresses used to establish the SCTP connection.

Reserved: Set to zero on transmission and ignored on receipt.

Exp: For experimental use [RFC3692].

SCTP Connection ID: An IPv4 or IPv6 address used to establish the SCTP connection. This field is omitted (length 0) for the Connection ID AFI 0.

Interface ID: An Interface ID is used to associate the connection a Join/Prune message is received over with an interface which is added or removed from an oif-list. When unnumbered interfaces are used or when a single Transport connection is used for sending and receiving Join/Prune messages over multiple interfaces, the Interface ID is used convey the interface from Join/Prune message sender to Join/Prune message receiver. When a PIM router sets a locally generated value for the Interface ID in the Hello TLV, it must send the same Interface ID value in all Join/Prune messages it is sending to the PIM neighbor.

4. Establishing Transport Connections

While a router interface is PORT enabled, a PIM-over-TCP or a PIM-over-SCTP option is included in the PIM Hello messages sent on that interface. When a router on a PORT-enabled interface receives a Hello message containing a PIM-over-TCP/PIM-over-SCTP Option from a new neighbor, or an existing neighbor that did not previously include the option, it switches to PORT mode for that particular neighbor.

When a router switches to PORT mode for a neighbor, it stops sending and accepting Native Join/Prune messages for that neighbor. Any state from previous Native Join/Prune messages is left to expire as normal. It will also attempt to establish a Transport connection (TCP or SCTP) with the neighbor. If both the router and its neighbor have announced both PIM-over-TCP and PIM-over-SCTP options, SCTP MUST be used.

When the router is using TCP it will compare the TCP Connection ID it announced in the PIM-over-TCP Capable Option with the TCP Connection ID in the Hello received from the neighbor. The router with the lower Connection ID will do an active Transport open to the neighbor Connection ID. The router with the higher Connection ID will do a passive Transport open. An implementation may open connections only on-demand, in that case it may be that the neighbor with the higher Connection ID does the active open, see Section 4.3. Note that the source address of the active open must be the announced Connection ID.

When the router is using SCTP, the IP address comparison need not be done since the SCTP protocol can handle call collision.

If PORT is used both for IPv4 and IPv6, both IPv4 and IPv6 PIM Hello messages are sent, both containing PORT Hello options. If two neighbors announce the same transport (TCP or SCTP) and the same Connection ID in the IPv4 and IPv6 Hello messages, then only one connection is established and is shared. Otherwise, two connections are established and are used separately.

The PIM router that performs the active open initiates the connection with a locally generated source transport port number and a well-known destination transport port number. The PIM router that performs the passive open listens on the well-known local transport port number and does not qualify the remote transport port number. See Section 5 for well-known port number assignment for PORT.

When a Transport connection is established (or reestablished), the two routers MUST both send a full set of Join/Prune messages for state for which the other router is the upstream neighbor. This is

needed to ensure that the upstream neighbor has the correct state. When moving from Datagram mode, or when the connection has gone down, the router cannot be sure that all the previous Join/Prune state was received by the neighbor. Any state received while in Datagram mode that is not refreshed, will be left to expire.

When a Transport connection goes down, Join/Prune state that was sent over the Transport connection is still retained. The neighbor should not be considered down until the neighbor timer has expired. This allows routers to do a control-plane switchover without disrupting the network. If a Transport connection is reestablished before the neighbor timer expires, the previous state is intact and any new Join/Prune messages sent cause state to be created or removed (depending on if it was a Join or Prune). If the neighbor timer does expire, only the upstream router, that has oif-list state, to the expired downstream neighbor will need to clear state. A downstream router, when an upstream neighboring router has expired, will simply update the RPF for the corresponding state to a new neighbor where it would trigger Join/Prune messages like it would in [RFC4601]. It is required of a PIM router to clear its neighbor table for a neighbor who has timed out due to neighbor holdtime expiration.

Note that, a Join sent over a Transport connection will only be seen by the upstream router, and thus will not cause routers on the link that do not use PIM PORT with the upstream router to possibly delay the refresh of Join state for the same state. Similarly, a Prune sent over a Transport connection will only be seen by the upstream router, and will thus never cause routers on the link on the link that do not use PIM PORT with the upstream router, to send a Join to override this Prune.

Note also, that a datagram PIM Join/Prune message for a said (S,G) or (*,G) sent by some router on a link will not cause routers on the same link that use a Transport connection with the upstream router for that state, to suppress the refresh of that state to the upstream router (because they don't need to periodically refresh this state) or to send a Join to override a Prune (as the upstream router will only stop forwarding the traffic when all joined routers that use a Transport connection have explicitly sent a Prune for this state, as explained in Section 6).

4.1. TCP Connection Maintenance

TCP is designed to keep connections up indefinitely during a period of network disconnection. If a PIM-over-TCP router fails, the TCP connection may stay up until the neighbor actually reboots, and even then it may continue to stay up until you actually try to send the neighbor some information. This is particularly relevant to PIM,

since the flow of Join/Prune messages might be in only one direction, and the downstream neighbor might never get any indication via TCP that the other end of the connection is not really there.

Implementations SHOULD support the use of TCP Keep-Alives, see [RFC1122] section 4.2.3.6. We recommend the use of Keep-Alives to be optional, allowing network administrators to use it as needed. Note that Keep-Alives can be used by a peer, independently of whether the other peer supports it. With the use of Keep-Alives one can detect that a connection is not working without sending any TCP data.

Most applications using TCP want to detect when a neighbor is no longer there, so that the associated application state can be released. Also, one wants to clean up the TCP state, and not keep half-open connections around indefinitely. This is accomplished by using PIM Hellos and by not introducing an application-specific or new PIM keep-alive message. Therefore, when a GENID changes from a received PIM Hello message, and a TCP connection is established or attempting to be established, the local side will tear down the connection and attempt to reopen a new one for the new instance of the neighbor coming up. However, if the connection is shared by multiple interfaces and the GENID changes only for one of them, then there was not a full reboot and the connection is likely to still work. In that case, the router should just resend all Join/Prune state for that particular neighbor. This is similar to how state is refreshed when GENID changes for PIM in datagram mode.

There may be situations where a router ignores some joins or prunes. E.g. due to wrong RP information or receiving joins on an RPF interface. A router may try to cache such messages and apply them later if only a temporary error. It may however also ignore the message, and later change its GENID for that interface to make the neighbor resend all state, including any that may have been previously ignored. It is possible that one receives Join/Prune messages for an interface/link that is down. As long as the neighbor has not expired, we recommend processing those messages as usual. If they are ignored, then the router should change the GENID for that interface when it comes back up, in order to get a full update.

4.2. Moving from PORT to Datagram Mode

There may be situations where an administrator decides to stop using PORT. If PORT is disabled on a router interface, we start expiry timers with the respective neighbor holdtimes as the initial values. Similarly if we receive a Hello message without a PORT Capable option from a neighbor, we start expiry timers for all Join/Prune state we have for that particular neighbor. The Transport connection should be shut down as soon as there are no more PIM neighborships using it.

That is, for the connection we have associated local and remote Connection IDs. When there is no PIM neighbor with that particular remote connection ID on any interface where we announce the local connection ID, the connection should be shut down.

4.3. On-demand versus Pre-configured Connections

Transport connections could be established when they are needed or when a router interface to other PIM neighbors has come up. The advantage of on-demand Transport connection establishment is the reduction of router resources. Especially in the case where there is no need for n^2 connections on a network interface. The disadvantage is additional delay and queueing when a Join/Prune message needs to be sent and a Transport connection is not established yet.

If a router interface has become operational and PIM neighbors are learned from Hello messages, at that time, Transport connections may be established. The advantage is that a connection is ready to transport data by the time a Join/Prune message needs to be sent. The disadvantage is there can be more connections established than needed. This can occur when there is a small set of RPF neighbors for the active distribution trees compared to the total number of neighbors. Even when Transport connections are pre-established before they are needed, a connection can go down and an implementation will have to deal with an on-demand situation.

Note that for TCP, it is the router with the lower Connection ID that decides whether to open a connection immediately, or on-demand. The router with the higher Connection ID should only initiate a connection on-demand. That is, if it needs to send a Join/Prune message and there is no currently established connection.

Therefore, this specification recommends but does not mandate the use of on-demand Transport connection establishment.

4.4. Possible Hello Suppression Considerations

This specification indicates that a Transport connection cannot be established until a Hello message is received. One reason for this is to determine if the PIM neighbor supports this specification and the other is to determine the remote address to use to establish the Transport connection.

There are cases where it is desirable to suppress entirely the transmission of Hello messages. In this case, it is outside the scope of this document on how to determine if the PIM neighbor supports this specification as well as an out-of-band (outside of the PIM protocol) method to determine the remote address to establish the

Transport connection.

4.5. Avoiding a Pair of Connections between Neighbors

To ensure there are not two connections between a pair of PIM neighbors, the following set of rules must be followed. Let A and B be two PIM neighbors where A's Connection ID is numerically smaller than B's Connection ID, and each is known to the other as having a potential PIM adjacency relationship.

At node A:

- o If there is already an established TCP connection to B, on the PIM-over-TCP port, then A MUST NOT attempt to establish a new connection to B. Rather it uses the established connection to send Join/Prune messages to B. (This is independent of which node initiated the connection.)
- o If A has initiated a connection to B, but the connection is still in the process of being established, then A MUST refuse any connection on the PIM-over-TCP port from B.
- o At any time when A does not have a connection to B which is either established or in the process of being established, A MUST accept connections from B.

At node B:

- o If there is already an established TCP connection to A, on the PIM-over-TCP port, then B MUST NOT attempt to establish a new connection to A. Rather it uses the established connection to send Join/Prune messages to A. (This is independent of which node initiated the connection.)
- o If B has initiated a connection to A, but the connection is still in the process of being established, then if A initiates a connection too, B MUST accept the connection initiated by A and must release the connection which it (B) initiated.

5. Common Header Definition

It may be desirable for scaling purposes to allow Join/Prune messages from different PIM protocol instances to be sent over the same Transport connection. Also, it may be desirable to have a set of Join/Prune messages for one address-family sent over a Transport connection that is established over a different address-family network layer.

To be able to do this we need a common header that is inserted and parsed for each PIM Join/Prune message that is sent on a Transport connection. This common header will provide both record boundary and demux points when sending over a stream protocol like Transport.

Each Join/Prune message will have in front of it the following common header in Type/Length/Value format. And multiple different TLV types can be sent over the same Transport connection.

To make sure PIM Join/Prune messages are delivered as soon as the TCP transport layer receives the Join/Prune buffer, the TCP Push flag will be set in all outgoing Join/Prune messages sent over a TCP transport connection.

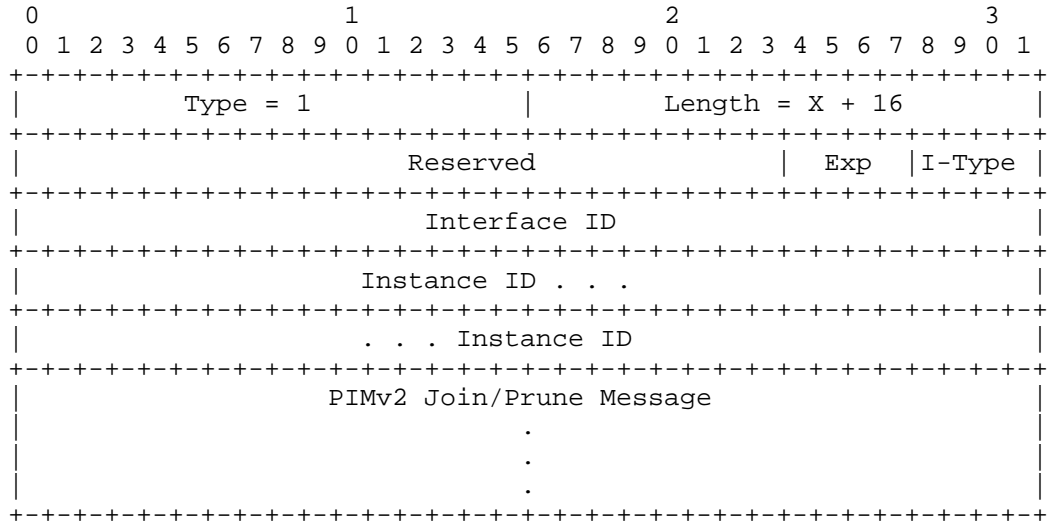
PIM messages will be sent using destination TCP port number 8471. When using SCTP as the reliable transport, destination port number 8471 will be used. See Section 10 for IANA considerations.

Join/Prune messages are error checked. This includes a bad PIM checksum, illegal type fields, illegal addresses or a truncated message. If any parsing errors occur in a Join/Prune message, it is skipped, and we proceed processing any following TLVs.

The TLV type field is 16 bits. The range 61440 - 65535 is for experimental use [RFC3692].

The current list of defined TLVs are:

IPv4 Join/Prune Message



The IPv4 Join/Prune common header is used when a Join/Prune message is sent that has all IPv4 encoded addresses in the PIM payload.

Length: In bytes for the value part of the Type/Length/Value encoding. Where X is the number of bytes that make up the PIMv2 Join/Prune message.

Reserved: Set to zero on transmission and ignored on receipt.

Exp: For experimental use [RFC3692].

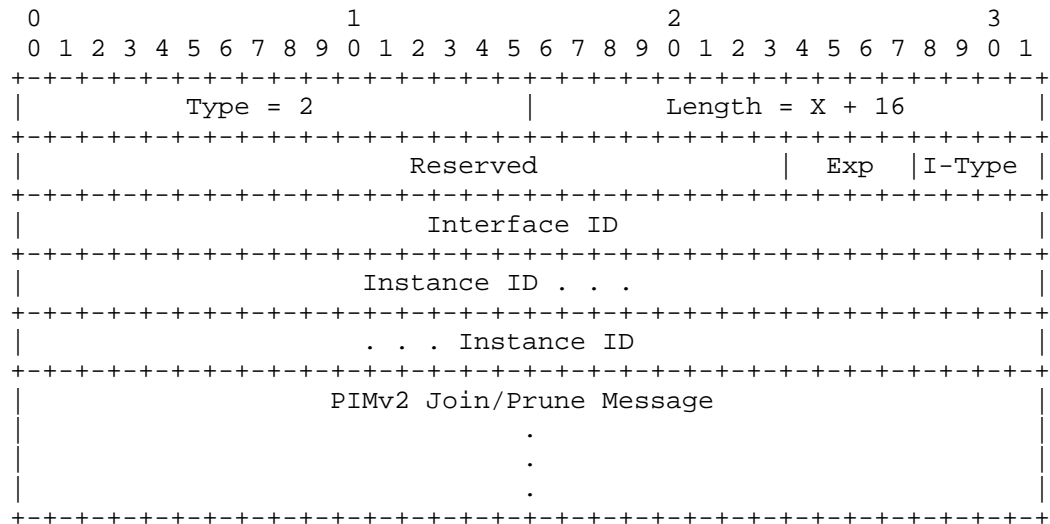
I-Type: Defines the encoding and semantics of the Instance ID field. Instance Type 0 means Instance ID is not used. Other values are not defined in this specification. A message with an unknown Instance Type MUST be ignored.

Interface ID: This is the Interface ID from the Hello TLV, defined in this specification, the PIM router is sending to the PIM neighbor. It indicates to the PIM neighbor what interface to associate the Join/Prune with.

Instance ID: This document only defines this for Instance Type 0. For type 0 the field should be set to zero on transmission and ignored on receipt. This field is always 64 bits.

PIMv2 Join/Prune Message: PIMv2 Join/Prune message and payload with no IP header in front of it. As you can see from the packet format diagram, multiple Join/Prune messages can go into one TCP/SCTP stream from the same or different Interface and Instance IDs.

IPv6 Join/Prune Message



The IPv6 Join/Prune common header is used when a Join/Prune message is sent that has all IPv6 encoded addresses in the PIM payload.

Length: In bytes for the value part of the Type/Length/Value encoding. Where X is the number of bytes that make up the PIMv2 Join/Prune message.

Reserved: Set to zero on transmission and ignored on receipt.

Exp: For experimental use [RFC3692].

I-Type: Defines the encoding and semantics of the Instance ID field. Instance Type 0 means Instance ID is not used. Other values are not defined in this specification.

Interface ID: This is the Interface ID from the Hello TLV, defined in this specification, the PIM router is sending to the PIM neighbor. It indicates to the PIM neighbor what interface to associate the Join/Prune with.

Instance ID: This document only defines this for Instance Type 0.
For type 0 the field should be set to zero on transmission and
ignored on receipt.

PIMv2 Join/Prune Message: PIMv2 Join/Prune message and payload with
no IP header in front of it. As you can see from the packet
format diagram, multiple Join/Prune messages can go into one TCP/
SCTP stream from the same or different Interface and Instance IDs.

6. Explicit Tracking

When explicit tracking is used, a router keeps track of join state for individual downstream neighbors on a given interface. This is done for all PORT joins and prunes. It may also be done for native join/prune messages, if all neighbors on the LAN have set the T bit of the LAN Prune Delay option. In the discussion below we will talk about ET (explicit tracking) neighbors, and non-ET neighbors. The set of ET neighbors always includes the PORT neighbors. The set of non-ET neighbors consists of all the non-PORT neighbors unless all neighbors have set the LAN Prune Delay T bit. Then the ET neighbors set contains all neighbors.

For some link-types, e.g. point-to-point, tracking neighbors is no different than tracking interfaces. It may also be possible for an implementation to treat different downstream neighbors as being on different logical interfaces, even if they are on the same physical link. Exactly how this is implemented and for which link types, is left to the implementer.

For (*,G) and (S,G) state, the router starts forwarding traffic on an interface when a Join is received from a neighbor on such an interface. When a non-ET neighbor sends a Prune, there is generally a small delay to see if another non-ET neighbor sends a Join to override the Prune. If there is no override, one should note that no non-ETP neighbor is interested. If no ET neighbors are interested, the interface can be removed from the oif-list. When a ET neighbor sends a Prune, one removes the join state for that neighbor. If no other ET or non-ET neighbors are interested, the interface can be removed from the oif-list. When a PORT neighbor sends a prune, there can be no Prune Override, since the Prune is not visible to other neighbors.

For (S,G,R) state, the router needs to track Prune state on the shared tree. It needs to know which ET neighbors have sent prunes, and whether any non-ET neighbors have sent prunes. Normally one would forward a packet from a source S to a group G out on an interface if a (*,G)-join is received, but no (S,G,R)-prune. With ET one needs to do this check per ET neighbor. That is, the packet should be forwarded unless all ET neighbors that have sent (*,G)-joins have also sent (S,G,R)-prunes, and if a non-ET neighbor has sent a (*,G)-join, whether there also is non-ET (S,G,R)-prune state.

7. Multiple Instances and Address-Family Support

Multiple instances of the PIM protocol may be used to support e.g. multiple address families. Multiple instances can cause a multiplier effect on the number of router resources consumed. To be able to have an option to use router resources more efficiently, muxing Join/Prune messages over fewer Transport connections can be performed.

There are two ways this can be accomplished, one using a common header format over a TCP connection and the other using multiple streams over a single SCTP connection.

Using the Common Header format described previously in this specification, using different TLVs, both IPv4 and IPv6 based Join/Prune messages can be encoded within a Transport connection. Likewise, within a TLV, multiple occurrences of Join/Prune messages can occur and are tagged with an instance-ID so multiple Join/Prune messages for different instances can use a single Transport connection.

When using SCTP multi-streaming, the common header is still used to convey instance information but an SCTP association is used, on a per-instance basis, to send data concurrently for multiple instances. When data is sent concurrently, head of line blocking, which can occur when using TCP, is avoided.

8. Miscellany

No changes expected in processing of other PIM messages like PIM Asserts, Grafts, Graft-Acks, Registers, and Register-Stops. This goes for BSR and Auto-RP type messages as well.

This extension is applicable only to PIM-SM, PIM-SSM and Bidir-PIM. It does not take requirements for PIM-DM into consideration.

9. Security Considerations

Transport connections can be authenticated using HMACs MD5 and SHA-1 similar to use in BGP [RFC4271] and MSDP [RFC3618].

When using SCTP as the transport protocol, [RFC4895] can be used, on a per SCTP association basis to authenticate PIM data.

10. IANA Considerations

This specification makes use of a TCP port number and a SCTP port number for the use of PIM-Over-Reliable-Transport that has been allocated by IANA. It also makes use of IANA PIM Hello Options allocations that should be made permanent. In addition, a registry for PORT message types is requested. The registry should cover the range 0 - 61439. An RFC is required for assignments in that range. This document defines two PORT message types. Type 1, IPv4 Join/Prune Message; and Type 2, IPv6 Join/Prune Message. The type range 61440 - 65535 is for experimental use [RFC3692].

11. Contributors

In addition to the persons listed as authors, significant contributions were provided by Apoorva Karan and Arjen Boers.

12. Acknowledgments

The authors would like to give a special thank you and appreciation to Nidhi Bhaskar for her initial design and early prototype of this idea.

Appreciation goes to Randall Stewart for his authoritative review and recommendation for using SCTP.

Thanks also goes to the following for their ideas and commentary review of this specification, Mike McBride, Toerless Eckert, Yiqun Cai, Albert Tian, Suresh Boddapati, Nataraj Batchu, Daniel Voce, John Zwiebel, Yakov Rekhter, Lenny Giuliano, Gorrry Fairhurst, Sameer Gulrajani, Thomas Morin and Dimitri Papadimitriou.

A special thank you goes to Eric Rosen for his very detailed review and commentary. Many of his comments are reflected as text in this specification.

13. References

13.1. Normative References

- [RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.
- [RFC1122] Braden, R., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, October 1989.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3618] Fenner, B. and D. Meyer, "Multicast Source Discovery Protocol (MSDP)", RFC 3618, October 2003.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC4895] Tuexen, M., Stewart, R., Lei, P., and E. Rescorla, "Authenticated Chunks for the Stream Control Transmission Protocol (SCTP)", RFC 4895, August 2007.
- [RFC4960] Stewart, R., "Stream Control Transmission Protocol", RFC 4960, September 2007.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

13.2. Informative References

- [AFI] IANA, "Address Family Indicators (AFIs)", ADDRESS FAMILY NUMBERS <http://www.iana.org/numbers.html>, February 2007.
- [HELLO-OPT] IANA, "PIM Hello Options", PIM-HELLO-OPTIONS per RFC4601 <http://www.iana.org/assignments/pim-hello-options>, March 2007.
- [RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.

Authors' Addresses

Dino Farinacci
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: dino@cisco.com

IJsbrand Wijnands
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: ice@cisco.com

Stig Venaas
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: stig@cisco.com

Maria Napierala
AT&T Labs
200 Laurel Drive
Middletown, New Jersey 07748>
USA

Email: mnapierala@att.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 28, 2011

S. Venaas
cisco Systems
October 25, 2010

A Registry for PIM Message Types
draft-ietf-pim-registry-01.txt

Abstract

This document provides instructions to IANA for the creation of a registry for PIM message types. It specifies initial content of the registry based on existing RFCs specifying PIM message types. It also specifies a procedure for registering new types.

In addition to this, one message type is reserved, and may be used for a future extension of the message type space.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 28, 2011.

Copyright Notice

Copyright (c) 2010 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
2. Security Considerations	4
3. IANA Considerations	5
3.1. Initial registry	5
3.2. Assignment of new message types	5
4. Acknowledgements	6
5. References	7
5.1. Normative References	7
5.2. Informative References	7
Author's Address	8

1. Introduction

Apart from this document, there is no existing document specifying a registry for PIM message types. PIM version 1 made use of IGMP [RFC1112] and there is an IGMP registry [IGMPREG] listing the message types used by PIM version 1. PIM version 2 however is not based on IGMP, and a separate PIM message type registry is needed. There are currently several RFCs specifying new PIM version 2 message types that should be in this new registry. They are the RFCs for PIM Dense Mode [RFC3973], PIM Sparse Mode [RFC4601] and Bidirectional PIM [RFC5015].

This document specifies the initial content of the new PIM message type registry based on those existing RFCs. This document also specifies a procedure for registering new PIM message types.

In addition to this, this document reserves one message type. This type may be used for a future extension of the message type space. The current message type space is only 4 bits, so it is not unlikely that this will be needed. How exactly the extension should be done is left to a future document.

2. Security Considerations

This document only creates an IANA registry. There may be a security benefit in a well-known place for finding information on which PIM message types are valid and how they are used. Apart from that there are no security considerations.

3. IANA Considerations

This document requests IANA to create a PIM message type registry. This should be placed in the "Protocol Independent Multicast (PIM)" branch of the tree. Each entry in the registry consists of message type, message name and references to the documents defining the type.

3.1. Initial registry

The initial content of the registry should be as follows.

Type	Name	Reference
0	Hello	[RFC3973] [RFC4601]
1	Register	[RFC4601]
2	Register Stop	[RFC4601]
3	Join/Prune	[RFC3973] [RFC4601]
4	Bootstrap	[RFC4601]
5	Assert	[RFC3973] [RFC4601]
6	Graft	[RFC3973]
7	Graft-Ack	[RFC3973]
8	Candidate RP Advertisement	[RFC4601]
9	State Refresh	[RFC3973]
10	DF Election	[RFC5015]
15	Reserved (for extension of type space)	[this document]

3.2. Assignment of new message types

Assignment of new message types is done according to the "IETF Review" model, see [RFC5226].

4. Acknowledgements

Thanks to Toerless Eckert for his suggestion to reserve a type for future message type space extension.

5. References

5.1. Normative References

- [RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, January 2005.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

5.2. Informative References

- [IGMPREG] IANA, "IGMP Type Numbers", IGMP TYPE NUMBERS - per RFC3228, BCP57 <http://www.iana.org/assignments/igmp-type-numbers>, June 2005.
- [RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, August 1989.

Author's Address

Stig Venaas
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: stig@cisco.com

Network working group
Internet Draft
Category: Standard Track
Created: October 18, 2010
Expires: April 2011

H. Liu
L. Zheng
T. Bai
Y. Yu
Huawei Technologies.

Single Stream Multicast Fast ReRoute (SMFRR) Method
draft-liu-pim-single-stream-multicast-frr-01

Abstract

This document proposes an IP multicast fast reroute method based on differentiating primary and backup PIM join. The multicast stream is only sent along one of the multicast primary and backup path, which enables the efficient multicast delivery under both normal and abnormal conditions.

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].

Status of this Memo

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

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

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 August 15, 2009.

Copyright Notice

Copyright (c) 2010 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.

Table of Contents

1. Introduction.....	3
2. Principle of Single Stream Solution.....	3
2.1. Primary and Backup Path Setup.....	3
2.2. Fault Processing.....	4
2.3. Fault Recovery.....	4
3. The Definition of packet format.....	5
3.1. Multicast FRR join Attribute.....	5
3.2. PIM multicast FRR Hello Options.....	5
4. Single Stream Implementation Options.....	6
4.1. Disabling all nodes on backup path.....	6
4.2. Disabling only root node on backup path.....	8
5. Security Considerations.....	8
6. References.....	8
6.1. Normative References.....	8
Authors' Addresses.....	9

1. Introduction

This document proposes an IP multicast fast reroute method based on differentiating primary and backup PIM join, which is called Single Stream multicast FRR. In this method, two multicast forwarding paths are established respectively by PIM primary join and backup join. Under normal conditions, only primary path is used to make the multicast data delivery. If the node or link on the primary path fails, the multicast data forwarding is switched to the backup path.

Because either primary or backup nodes forward multicast data packets, they should be able to identify on which path they are located to make appropriate forwarding decision. One feasible solution is to include a new join attribute in a PIM backup join message to set up backup multicast path whose entries are disabled by default. If a failure is detected on the primary path, the backup nodes are notified and the entries which were previously disabled are enabled for multicast data forwarding.

The Single stream FRR solution has the advantages of implementing fast multicast protection and of avoiding double multicast bandwidth occupation in both normal and abnormal situations.

2. Principle of Single Stream Solution

2.1. Primary and Backup Path Setup

The backup multicast path is set up using backup PIM join. The join is sent by the initiating node (i.e. the downstream converge point of primary and backup paths) from a backup IP FRR upstream interface or from a statically configured backup interface towards the multicast source. The join is transmitted hop-by-hop upwards and is terminated when reaching the root of the multicast tree (i.e. Source DR or RP), or when merging with primary forwarding states created by primary join. On the merging point, only the primary states are maintained.

The forwarding state(s) on backup path are disabled by default for data forwarding when being created by the backup joins, which requires the backup join to be flagged to be differentiated from the primary ones. A new join attribute [RFC5384] (referred to as e.g. Multicast FRR join Attribute, or MFA), is suggested to be introduced to serve this purpose and a new hello option for this attribute should be defined to negotiate this capability. The format of the

attribute and its hello option are respectively defined in section 3.1 and 3.2

The establishing of primary path could be a normal PIM join process. In this case an ordinary PIM join is generated on the initiating node of primary path and is sent hop-by-hop upstream until the join arrives at the root of the tree or at the other valid forwarding branch.

2.2. Fault Processing

The fault on the primary path could be detected by using some fault detection mechanism (e.g. BFD protocol), which is configured to be run between each pair of PIM neighbors. If error condition occurs, the node on the upstream or downstream of the error point will possibly detect it and should pass this error condition to the backup path, and enable multicast data forwarding on it.

As the node on the primary path detects a failure, it could choose to flood the failure notification packet to all its PIM neighbors until all the PIM routers in the area get the notification. To prevent excessive transmission of these packets, the sending and forwarding of the packets should be rate-limited. There are other options such as setting up special fault notification tree with reserved multicast address and etc.

After the enabling of the backup path triggered by the fault notification, the multicast data will be forwarded along the backup path to the initiating node of the backup path. The initiating point will change the backup incoming interface (IIF) as its RPF interface if no data is available from the primary IIF.

2.3. Fault Recovery

If primary path heals, multicast forwarding could choose to switch back to the primary path. Once the data is received from the primary IIF, the initiating node will change its RPF interface to its primary IIF. The node may also send a PIM prune message to tear down the backup path, and may possibly after waiting for a specified period of time, re-setup the backup path without stream using the same process as described in section 2.1.

3. The Definition of packet format

3.1. Multicast FRR join Attribute

The format of the join attribute is defined as:

```

+-----+-----+-----+-----+-----+-----+-----+-----+
|F|E| Attr_Type |      Length      |      Flags      | Path Count |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Path ID                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
~                                     . . .                                     ~
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Path ID                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

- F-bit, Transitive Attribute. If this bit is set, the attribute is a transitive attribute; otherwise, it is a non-transitive attribute [RFC5384].

- E-bit, End of Attributes. If this bit is set, then this is the last Join Attribute appearing in the Encoded-Source Address field specified by [RFC5384].

- Attr_Type, Type of the Attribute. It should be set to a new value (e.g.) for this MFA join attribute, e.g., taking value of 8.

- Length, a 1-octet field specifying the length in octets, encoded as an unsigned binary integer, of the value field.

- Flags, flags for the methods of setting up of primary or backup paths. For the rightmost bit, 0 is for a primary join, 1 for backup join. Other bits are reserved for the future definition.

- Path Count, the number of Path ID.

- Path ID, the Identification for this path. It may be an interface ID or a logical number to identify a primary path.

3.2. PIM multicast FRR Hello Options

This multicast FRR Hello options are used for the PIM neighbors to negotiate the capability of multicast FRR join attribute. It has the format prescribed in [RFC5384] and the OptionType is defined a new value representing this MFA attribute.


```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|           OptionType           |           OptionLength           |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|                               |                               |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
- OptionType = 38

- OptionLength = 8

- OptionValue, reserved for future use

```

4. Single Stream Implementation Options

4.1. Disabling all nodes on backup path

In this method, when backup join is transmitted to set up the backup path, the forwarding states of all backup nodes are by default disabled for multicast data forwarding when being created. When backup join arrives at a primary node that has primary forwarding state, it is "absorbed" and will not create any backup state there. Because each node on the backup path could be disabled or enabled for data forwarding, it is possible to implement relatively precise control of path switching.

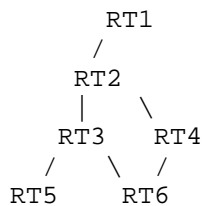


Figure 1

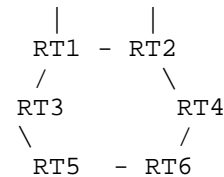


Figure 2

Figure 1 is an example of an arbitrary tree topology. Supposing RT6 has a downstream receiver and it is the initiating node of both the primary and backup path for this receiver. RT2-RT3-RT6 is setup as the primary path by primary join, and RT2-RT4-RT6 as the backup path by backup join. The backup forwarding entries for the backup path, i.e. the outgoing interfaces of RT2 (the one towards RT4) and RT4 (towards RT6), are all disabled for multicast forwarding. Only primary path imports multicast stream through RT2 to RT6 and to the receiver.

If link between RT3 and RT6 goes down, the failure will be detected by RT6. The fault notification will be generated by RT6 and be

notified to RT4 and RT2 on backup path, by flooding or through fault multicast tree which is pre-established on RT4 and RT2 by back join with RT6 as the source of the tree. The nodes RT4 and RT2 will be enabled the data forwarding on their outgoing interfaces, and the data will be imported from RT2, through RT4, to RT6 and the receiver.

In the ring topology shown in figure 2, supposing RT3 has a receiver downstream, the primary path for it is RT1-RT3 and takes the duty of data forwarding. The backup path is RT2-RT4-RT6-RT5-RT3 and the backup outgoing interface on each of them is disabled when the forwarding state is created. If node RT1 undergoes failure, it will be detected by RT3 and be notified by flooding or by multicast fault tree which is pre-established on RT5, RT6, RT4, and RT2, with RT3 as the source of the tree. After enabling data forwarding for these nodes, the traffic will be delivered along backup path to RT3 and to the receiver. Each node on the ring processes in the similar manner, if it has downstream multicast receiver. If any upstream failure on the primary path occurs, the node will turn to receive reverse stream from the backup path.

Because a backup node or path might provide protection for more than one primary path, the identification of the primary path should be bound to its own backup multicast entries, which requires the identification to be carried in the backup join during setting up of backup path, and in the fault notification to enable the forwarding of these entries.

In normal cases, a primary path is identified by the primary IIF ID of an initiating node. In figure 1, this ID is the upstream interface ID of RT6 towards RT3. Its correlation with backup forwarding entries are maintained at RT4 and RT2. The backup path RT2-RT4-RT6 is used to protect failure detected by RT6 (i.e. RT3 node failure or RT3-RT6 link failure). To provide protection for the whole primary network, each primary node is required to have a backup interface to form disjoint backup path for the upstream node/link to be protected, which is generally the case for ring topology and dual-homing protection tree topology.

As an extension, the primary ID could also be the collection of all interface IDs of a primary path (i.e. upstream interface IDs of RT3 and RT6 in figure 1), which could be configured on the initiating point (i.e. RT6) and be carried in backup join. The fault notification still carries the interface ID of downstream detecting node. Because the backup path is set up according to the interface ID collection for the whole primary path, one backup path can

provide protection for a complete primary path (i.e. RT2-RT3-RT6), rather than for only one hop distance in the former case.

4.2. Disabling only root node on backup path

In this method, when backup join is sent to setup the backup path, only the root node is disabled of its multicast data forwarding. The forwarding states on other nodes on the backup path are kept normal. In normal condition, the only stream comes from the primary path established by the primary join. If error occurs on the primary path, the root node of the backup path is notified of the failure, it then enables its data forwarding and the data stream will be delivered from the backup path to the receiver.

Because only the ingress node of the backup path is disabled, the method requires the backup path not to intersect with the primary path for the intermediate nodes and can be applied to multiple tree topologies. E.g., the primary join and backup join can be used to setup primary and backup trees with only primary tree makes the multicast forwarding in normal condition. When failure occurs on the primary tree, the root node of the backup tree could be notified to open its data forwarding and the multicast data will delivered over the backup tree to the receiver.

5. Security Considerations

They will be described in the later version of this draft.

6. References

6.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to indicate requirement levels", RFC 2119, March 1997.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC5384] A. Boers, I. Wijnands, E. Rosen, "The Protocol Independent Multicast (PIM) Join Attribute Format", RFC 5384, November 2008

[RFC5880] Katz, D., and Ward, D., "Bidirectional Forwarding Detection", RFC 5880, June, 2010.

Authors' Addresses

Hui Liu
Huawei Technologies Co., Ltd.
Huawei Bld., No.3 Xinxu Rd.
Shang-Di Information Industry Base
Hai-Dian Distinct, Beijing 100085
China

EMail: Liuhui47967@huawei.com

Lianshu Zheng
Huawei Technologies Co., Ltd.
Huawei Bld., No.3 Xinxu Rd.
Shang-Di Information Industry Base
Hai-Dian Distinct, Beijing 100085
China

EMail: verozheng@huawei.com

Tao Bai
Huawei Technologies Co., Ltd.
No.156 BeiQing Rd.
Hai-Dian Distinct, Beijing 100094

EMail: baitao_bys@huawei.com

YunFu Yu
Huawei Technologies Co., Ltd.
No.156 BeiQing Rd.
Hai-Dian Distinct, Beijing 100094
China

EMail: yuyunfu@huawei.com

