PIM Working Group
Internet-Draft
Expires: October 28, 2010

B. Joshi
Infosys Technologies Ltd.
A. Kessler
Cisco Systems, Inc.
D. McWalter
Metaswitch Networks
April 26, 2010

PIM Group-to-RP Mapping
draft-ietf-pim-group-rp-mapping-04.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 October 28, 2010.

Copyright Notice

Table of Contents

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. 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 in 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:

- It does not consider the origin of a Group-to-RP mapping and therefore will treat all of them equally.

- 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:

- 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.

- 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.

- 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

- A Group-to-RP mapping learned for PIM-BIDIR mode is preferred to an entry learned for PIM-SM mode.

- Dynamic group-to-RP mapping mechanisms are filtered at domain boundaries or for policy enforcement inside a domain.

5. Common use cases

- 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 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.

- 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.

- 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.

- 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 'pimStaticRPPrecedence' 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 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

[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
Metaswitch Networks
100 Church Street
Enfield EN2 6BQ
UK

Email: dmcw@metaswitch.com