PIM Working Group
Internet-Draft
Updates: 4601 (if approved)
Intended status: Standards Track
Expires: August 1, 2011

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

January 28, 2011

PIM Group-to-RP Mapping
draft-ietf-pim-group-rp-mapping-10.txt

Abstract

Each PIM-SM router in a Protocol Independent Multicast (PIM) Domain which supports Any Source Multicast (ASM) maintains Group-to-RP mappings which are used to identify a Rendezvous Point(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

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 as described in section 4.

It is critical that each router select the same 'RP' for a specific multicast group address otherwise full multicast connectivity will not be established. 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] includes a number of objects to allow an administrator to set the precedence for Group-to-RP mappings which are learned statically or dynamically and stored in the 'pimGroupMappingTable'. The Management Information Base (MIB) module also defines an algorithm that can be applied to the data contained in the 'pimGroupMappingTable' to determine Group-to-RP mappings. However, this algorithm is not completely deterministic, because it includes an implementation-specific 'precedence' value.

Network Management stations will be able to deduce which RPs will be selected by applying the algorithm from this document to the list of Group-to-RP mappings from the 'pimGroupMappingTable'. The algorithm provides MIB visibility into how routers will apply Group-to-RP mappings and also fixes the protocol inconsistency with how different vendors select the Group-to-RP mappings to create multicast forwarding state.

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

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]. 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 Bidirectional (BIDIR) Mode.

o Dynamic group-to-RP mapping mechanisms

The term Dynamic group-to-RP mapping mechanisms in this document refers to Bootstrap Router (BSR) [RFC5059] 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

The 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.

The algorithm defined in this document updates algorithm defined in PIM-SM (Section 4.7.1 in [RFC4601]). The new algorithm is backward compatible and will produce the same result only if the Group-to-RP mappings are learned from a single mapping source. The full benefits of the new algorithm will not be realized until it is widely deployed.

4. Assumptions

We have made following assumptions in defining this algorithm:

- 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 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 learned for PIM-BIDIR mode is preferred to an entry learned for PIM-SM mode as stipulated by section 3.3 of [RFC5059].

- 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. This is not an issue for IPv6 Multicast address ranges.

- 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 MIB objects.

6. Proposed algorithm

The following algorithm deterministically chooses between several Group-to-RP mappings for a specific group. It also addresses the above mentioned shortcomings in the existing mechanism.

1. If the Multicast Group Address being looked up contains an embedded RP, the RP address extracted from the Group address is selected as the Group-to-RP mapping.

2. If the Multicast Group Address being looked up is in the Source Specific Multicast (SSM) range or is configured for Dense mode, no Group-to-RP mapping is selected, and this algorithm terminates. The fact that no Group-to-RP mapping has been selected can be represented in the PIM-STD-MIB module by setting the address type of the RP to 'unknown' as described in Section 8.

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, is selected.

4. If there are no entries available, then the Group-to-RP mapping is undefined and this algorithm terminates.

5. A longest prefix match is performed on the subset of Group-to-RP Mappings.

   * If there is only one entry available then that entry is selected as the 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 entry is selected as the 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 is preferred.

   * If there is only one entry available then that entry is selected as the 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 as defined in section 4.7.2 of [RFC4601] will be used to choose the RP. The RP with the highest resulting hash value will be selected. Please look at section 10 for consideration of hash for BIDIR-PIM and BSR.

   * 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 (numerically greater). This will serve as a final tiebreaker.

7. Interpretation of MIB Objects

As described in [RFC5060] the Group-to-RP mapping information is summarized in the pimGroupMappingTable. The precedence value is stored in the 'pimGroupMappingPrecedence' object which covers both the dynamically learned Group-to-RP mapping information as well as the static configuration. For static configurations, the 'pimGroupMappingPrecedence' object uses the value of the 'pimStaticRPPrecedence' object from the pimStaticRPTable table.

The algorithm defined in this document does not use the concept of precedence and therefore the values configured in the 'pimGroupMappingPrecedence' and 'pimStaticRPPrecedence' objects in the PIM-STD-MIB module [RFC5060] are not applicable to the new algorithm. The objects still retain their meaning for 'legacy' implementations, but since the algorithm defined in this document is to be used in preference to that found in PIM-SM [RFC4601] and PIM-STD-MIB [RFC5060], the values of these objects will be ignored on implementations that support the new algorithm.

8. Clarification for MIB Objects

An implementation of this specification can continue to be managed using the PIM-STD-MIB [RFC5060]. When a Group-to-RP mapping entry is created in the pimGroupMappingTable with RP address type in the pimGroupMappingRPAddressType object is set to unknown(0), and the PIM Mode in the pimGroupMappingPimMode object is set to either ssm(2) or dm(5) to represent group ranges for SSM or Dense mode.

Also, all the entries which are already included in the SSM Range table in the IP Mcast MIB [RFC5132] are copied to the pimGroupMappingTable. Such entries have their type in the pimGroupMappingOrigin object set to configSsm(3), and the RP address type in the pimGroupMappingRPAddressType object set to unknown(0) as described above.

9. Use of dynamic Group-to-RP mapping protocols

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.

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 on all routers in the domain. This can be important at domain border routers, and is likely to avoid conflicts under misconfiguration (when routers receive overlapping sets of Group-to-RP mappings) and when configuration is changing.

An implementation of PIM that supports only one mechanism for learning Group-to-RP mappings MUST 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 filter 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 filter out 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 enhances an existing algorithm to deterministically choose between several Group-to-RP Mappings for a specific group. Different routers may select a different Group-to-RP Mapping for the same group if the Group-to-RP Mappings learned in these routers are not consistent. For example: let us assume that BSR is not enabled in one of the routers and so it does not learn any Group-to-RP Mappings from BSR. Now the Group-to-RP Mappings learned in this router may not be consistent with other routers in the network, it may select a different RP or may not select any RP for a given group. Such situations can be avoided if the mechanisms used to learn Group-to-RP Mappings are secure and consistent across the network. Secure transport of the mapping protocols can be accomplished by using authentication with IPsec as described in section 6.3 of [RFC4601].

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.

[RFC5132] McWalter, D., Thaler, D., and A. Kessler, "IP Multicast MIB", RFC 5132, December 2007.

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