MBONED Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 10, 2011

H. Asaeda
Y. Uchida
Keio University

March 9, 2011

IGMP/MLD-Based Explicit Membership Tracking Function for Multicast Routers
draft-asaeda-mboned-explicit-tracking-02

Abstract

This document describes the IGMP/MLD-based explicit membership tracking function for multicast routers. The explicit tracking function is useful for accounting and contributes to saving network resource and fast leaves (i.e. shortened leave latency).

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 September 10, 2011.

Table of Contents

1. Introduction

The Internet Group Management Protocol (IGMP) [2] for IPv4 and the Multicast Listener Discovery Protocol (MLD) [3] for IPv6 are the standard protocols used by listener hosts and multicast routers. When a host starts listening particular multicast channels, it sends IGMP/MLD State-Change Report messages specifying the corresponding channel information as the join/leave request to its upstream router (i.e., an adjacent multicast router or IGMP/MLD proxy [8]). This "unsolicited" Report is sent only once upon reception.

IGMP/MLD are non-reliable protocols; the unsolicited Report messages may be lost or not be reached to upstream routers. To recover the problem, the routers need to update membership information by sending IGMP/MLD General Query messages periodically. Member hosts then reply with "solicited" Report messages whenever they receive the Query messages.

Multicast routers are able to periodically maintain the multicast listener (or membership) state of downstream hosts attached on the same link by getting unsolicited Report messages and synchronize the actual membership state within the General Query timer interval (i.e., [Query Interval] value defined in [2][3].) However, this approach does not guarantee that the membership state is always perfectly synchronized. To minimize the possibility of having the outdated membership information, routers may shorten the periodic General Query timer interval. Unfortunately, this would increase the number of transmitted solicited Report messages and induce network congestion. And the more the network congestion is occured, the more IGMP/MLD Report messages may be lost and the membership state information may be outdated in the router.

The IGMPv3 [2] and MLDv2 [3] protocols can provide the capability of keeping track of downstream (adjacent) multicast listener state to multicast routers. This document describes the "IGMP/MLD-based explicit member tracking function" for multicast routers and details the way for routers to implement the function. By enabling the explicit tracking function, routers can keep track of the downstream multicast membership state. This function implements the following requirements:

- Per-host accounting
- Reducing the number of transmitted Query and Report messages
- Shortening leave latencies

- Maintaining multicast channel characteristics (or statistics)

where this document mainly focuses on the above second and third bullets in the following sections.

The explicit tracking function does not change message formats used by the standard IGMPv3 [2] and MLDv2 [3], and their lightweight version protocols [4]. It does not change a multicast data sender's and receiver's behavior as well.

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 [1].

3. Explicit Tracking Function

3.1. Reducing the Number of Specific Queries

The explicit tracking function reduces the number of Group-Specific or Group-and-Source Specific Query messages transmitted from a router, and then the number of Current-State Report messages transmitted from member hosts. As the result, network resources used for IGMP/MLD query-and-reply communications between a router and member hosts can be saved.

According to [2] and [3], whenever a router receives the State-Change Report, it sends the corresponding Group-Specific or Group-and-Source Specific Query messages to confirm whether the Report sender is the last member host or not. After getting these Query messages, all member hosts joining the corresponding channel reply with own Current-State Report messages. This condition requires transmitting a number of Current-State Report messages and consumes network resources especially when many hosts have been joining the same channel.

On the other hand, if a router enables the explicit tracking function, it does not need to always ask Current-State Report message transmission to the member hosts whenever it receives the State-Change Report. This is because the explicit tracking function works with the expectation that the State-Change Report sender is the last remaining member of the channel. Even if this expectation is wrong (i.e., the State-Change Report sender was not the sole member), other members remaining in the same channel will reply with identical Report messages, so the end result is the same and no problem occurs. (Section 4 details the point.)

In addition, the processing of IGMP membership or MLD listener reports consumes CPU ressources on the IGMP/MLD querier devices itself. Therefore, the explicit tracking function reduces not only the network load but also the CPU load on the querier devices as well.

3.2. Shortening Leave Latencies

The explicit tracking function works with the expectation that the State-Change Report sender is the last remaining member of the channel. Thanks to this functionality, a router can tune timers and values related to decide that the State-Change Report sender was the sole member.

The [Last Member Query Interval] (LMQI) and [Last Listener Query Interval] (LLQI) values specify the maximum time allowed before

sending a responding Report. The [Last Member Query Count] (LMQC) and [Last Listener Query Count] (LLQC) are the number of Group-Specific Queries or Group-and-Source Specific Queries sent before the router assumes there are no local members. The [Last Member Query Time] (LMQT) and [Last Listener Query Time] (LLQT) values are the total time the router should wait for a report, after the Querier has sent the first query.

The default values for LMQI/LLQI defined in the standard specifications [2][3] are 1 second. For the router enabling the explicit tracking function, LMQI/LLQI SHOULD be 1 second or shorter. The LMQC/LLQC MAY be set to "1" for the router, whereas their default values are the [Robustness Variable] value whose default value is "2". Smaller LMQC/LLQC give smaller LMQT/LLQT; this condition shortens the leave latencies.

3.3. Considerations

As with the basic concepts of IGMP and MLD, the explicit tracking function does not guarantee the membership state is always perfectly synchronized; routers enabling the explicit tracking function still need to send IGMPv3/MLDv2 Query messages and inquire solicited IGMPv3/MLDv2 Report messages from downstream members to maintain downstream membership state.

- IGMP/MLD messages are non-reliable and may be lost in the transmission, therefore routers need to confirm the membership by sending Query messages.

- To preserve compatibility with older versions of IGMP/MLD, routers need to support downstream hosts that are not upgraded to the latest versions of IGMP/MLD and run the report suppression mechanism.

- It is impossible to identify hosts when hosts send IGMP reports with a source address of 0.0.0.0.

Regarding the last bullet, the IGMPv3 specification [2] mentions that an IGMPv3 Report is usually sent with a valid IP source address, although it permits that a host uses the 0.0.0.0 source address (as it happens that the host has not yet acquired an IP address), and routers MUST accept a report with a source address of 0.0.0.0. The MLDv2 specification [3] mentions that an MLDv2 Report MUST be sent with a valid IPv6 link-local source address, although an MLDv2 Report can be sent with the unspecified address (::), if the sending interface has not acquired a valid link-local address yet. [3] also mentions that routers silently discard a message that is not sent with a valid link-local address or sent with the unspecified address,

without taking any action, because of the security consideration.

Another concern is that the explicit tracking function requires additional processing capability and a possibly large memory for routers to keep all membership states. Especially when a router needs to maintain a large number of member hosts, this resource requirement may be potentially-impacted. Operators may decide to disable this function when their routers do not have enough memory resources.

4. Membership State Information

The explicit tracking function is implemented with the following membership state information:

(S, G, number of receivers, (receiver records))

where each receiver record is of the form:

(IGMP/MLD Membership/Listener Report sender's address)

This state information must work properly when a receiver (i.e., Report sender) sends the same Report messages multiple times.

In the state information, each "S" and "G" indicates a single IPv4/ IPv6 address. "S" is set to "Null" for an Any-Source Multicast (ASM) communication (i.e., (*,G) join reception). In order to simplify the implementation, the explicit tracking function does not keep the state of (S,G) join with EXCLUDE filter mode. If a router receives (S,G) join/leave request with EXCLUDE filter mode from the downstream hosts, it translates the join/leave request to (*,G) join state/leave requst and records the state and the receivers' addresses into the maintained membership state information. Note that this membership state translation does not change the routing protocol behavior; the routing protocol must deal with the original join/leave request and translate the request only for the membership state information.

5. Multicast Router Behavior

The explicit tracking function makes routers expect whether the State-Change Report sender is the last remaining member of the channel. Therefore the router transmits a corresponding Current-State Report message only when the router thinks that the State-Change Report sender is the last remaining member of the channel. This contributes to saving the network resources and also shortening leave latency.

To synchronize the membership state information, when a multicast router receives a Current-State or State-Change Report message, it adds the receiver IP address to or delete from the receiver records or creates the corresponding membership state information. If there are no more receiver records left, the membership state information is deleted from the router.

However, the membership state information may be still outdated in the router. It may be happened especially in a mobile multicast environment that some member hosts have joined to or left from the network without sending State-Change Report messages. Or, some State-Change Report messages are lost due to network congestion. Therefore, the router enabling the explicit tracking function MUST send the periodic General Query regularly.

Regarding the leave latency, as specified in Section 3.2, the explicit tracking function contributes to the fast leave by setting LMQI/LLQI to "1" second or shorter and LMQC/LLQC to "1". However, if LMQC/LLQC is configured "2" or bigger value, then the router's behavior MAY be changed from the standard specification. According to [2] and [3], a router sends a Group- (and-Source) Specific Query [LMQC - 1] or [LLQC - 1] times when it receives State Change Report message (e.g. leave request) from a member host, in order to confirm whether or not the host is the only remaining member. However, this document RECOMMENDS that if the router enabling the explicit tracking function receives the corresponding Current State Report before the Specific Query retransmission, it cancels sending the same Specific Query for other [LMQC - 1] or [LLQC - 1] times.

Note that there is some risk that a router misses or looses Report messages sent from remaining members if the router adopts small LMQC/LLQC; however the wrong expectation would be lower happened for the router enabling the explicit tracking function. And to avoid the problem, a router can start sending a Group- (and-Source) Specific Query message when it expects the number of the remaining members is small, such as 5, but not 0.

6. Interoperability and Compatibility

The explicit tracking function does not work with the older versions of IGMP or MLD, IGMPv1 [5], IGMPv2 [6] or MLDv1 [7], because a member host using these protocols adopts a report suppression mechanism by which a host would cancel sending a pending membership Reports if a similar Report was observed from another member on the network.

If a multicast router enabling the explicit tracking function changes its compatibility mode to the older versions of IGMP or MLD, the router SHOULD turn off the explicit tracking function but SHOULD NOT flush the maintained membership state information (i.e., keep the current membership state information as is). When the router changes back to IGMPv3 or MLDv2 mode, it SHOULD resume the function with the kept membership state information, even if the state information is outdated. This manner would give "smooth state transition" that does not initiate the membership state from scratch and synchronizes the actual membership state smoothly.

There are several points TBD in the further discussions regarding the interoperability and compatibility issues. At first, it is necessary whether a multicast router enabling the explicit tracking function needs to detect adjacent routers that do not support the explicit tracking function on the link or not. After the clarification, this document will describe the method how to detect them. It would be done by a new signaling message, but the new message leads compatibility problems for older routers or other routing protocols such as PIM-DM. All of these discussions are TBD.

7. Security Considerations

TBD.

8. Acknowledgements

Toerless Eckert, Nicolai Leymann, Stig Venaas, and others provided many constructive and insightful comments.

9. References

9.1. Normative References

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

[2] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.

[3] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.

[4] Liu, H., Cao, W., and H. Asaeda, "Lightweight Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Version 2 (MLDv2) Protocols", RFC 5790, February 2010.

9.2. Informative References

[5] Deering, S., "Host Extensions for IP Multicasting", RFC 1112, August 1989.

[6] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2373, July 1997.

[7] Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, October 1999.

[8] Fenner, B., He, H., Haberman, B., and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.

Authors' Addresses

Hitoshi Asaeda
Keio University
Graduate School of Media and Governance
5322 Endo
Fujisawa, Kanagawa 252-0882
Japan

Email: asaeda@wide.ad.jp
URI: http://www.sfc.wide.ad.jp/~asaeda/

Yogo Uchida
Keio University
Graduate School of Media and Governance
5322 Endo
Fujisawa, Kanagawa 252-0882
Japan

Email: uchida@sfc.wide.ad.jp

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 19, 2011

S. Gulrajani
S. Venaas
cisco Systems

February 15, 2011

An Interface ID Hello Option for PIM
draft-gulrajani-pim-hello-intid-00.txt

Abstract

This document defines a new PIM Hello option to advertise an interface id that can be used by PIM protocols to uniquely identify an interface of a neighboring router.

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 August 19, 2011.

Copyright Notice

Table of Contents

1. Introduction

This document defines a new option for use in PIM Hello messages [RFC4601] to carry an Interface Identifier. A router generates identifiers for each of its PIM enabled interfaces so that each interface has a different identifier. The identifiers can optionally be generated so that they are unique within, e.g., an administrative domain.

An example where this Interface Identifier can be used is with PIM PORT [I-D.ietf-pim-port], where a single Transport connection is used between two routers that have multiple interfaces connecting them. If these interfaces have unnumbered or IPv6 Link local addresses, the Interface Identifier included in the PORT Join/Prune message will identify which interface the message is associated with. For PIM PORT the Router Identifier is not needed, and it can be set to zero.

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].

2. Interface Identifier Option

The Interface Identifier option is used to identify which interface of a neighboring router a PIM Hello [RFC4601] is sent on. This allows PIM protocols to refer to, or identify, a particular interface on a neighboring router.

The Interface Identifier option need only be included in PIM Hello messages if the router supports protocols that require it. An implementation MAY choose to always include it. How exactly the Interface Identifier is used, and the uniqueness requirements, is left to the specifications of the PIM protocols that make use of it. It is assumed that different protocols may have different minimum requirements for stability and uniqueness, but that they have no maximum requirement. When specified, these protocols should indicate what their minimum requirements are.

The Interface Identifier consists of 64 bits. The lower 32 bits form a Local Interface Identifier, and the high 32 bits a Router Identifier.

2.1. Local Interface Identifier

The 32 bit Local Interface Identifier is selected so that it is unique among the router's PIM enabled interfaces. That is, there MUST NOT be two PIM interfaces with the same Local Interface Identifier. While an interface is up, the Identifier MUST always be the same once it has been allocated. If an interface goes down and up, the router SHOULD use the same Identifier. Many systems makes use of an ifIndex [RFC1213], which can be used as a Local Interface Identifier.

The Local Interface Identifier MUST be non-zero. The reason for this, is that some protocols may want to only optionally refer to an Interface using the Interface Identifier Hello option, and use the value of 0 to show that it is not referred to. Note that the value of 0 is not a valid ifIndex as defined in [RFC1213].

2.2. Router Identifier

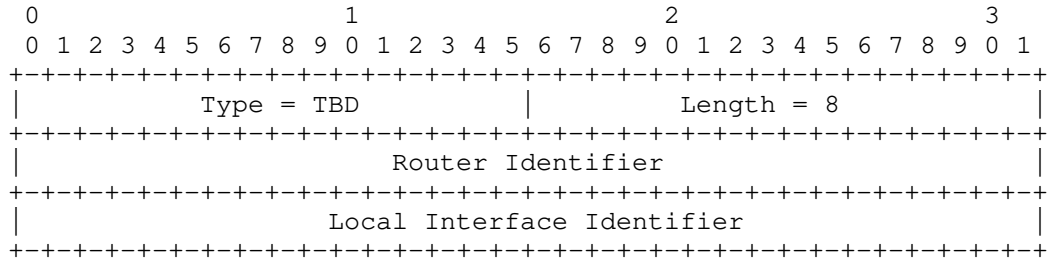
The 32 bit Router Identifier may be used to uniquely identify the router. It may be selected to be unique within some administrative domain, or possibly globally unique. In which scope it needs to be unique depends on the protocol utilizing it. Routers often have such an identifier derived from an IPv4 address or manual configuration. Protocols like BGP [RFC4271] and OSPFv2 [RFC2328] are other protocols making use of 32 bit identifiers for routers. One may use the same identifier to construct the Interface Identifier option, provided it

meets the stability and uniqueness requirements of protocols making use of this option.

The value 0 has a special meaning for the Router Identifier. It means that no Router Identifier is used. If a router only supports protocols that require the Interface Identifier to be unique for one router (only making use of the Local Interface Identifier), then the implementation MAY set the Router Identifier to zero.

3. Message Format

Option Type: Interface Identifier



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

Length: In bytes for the value part of the Type/Length/Value encoding. The Interface Identifier will be 8 bytes long.

Local Interface Identifier: The Local Interface Identifier is a 4 byte identifier that is unique among all PIM enabled interfaces on a router.

Router Identifier: The Router Identifier is a 4 byte identifier uniquely identifying the router within some scope. It MAY be 0 when no protocols require a Router Identifier.

4. Security Considerations

The Interface Identifier is included in PIM Hello Messages. Apart from the general security considerations for PIM messages, the only additional concern is what may happen if a spoofed PIM message is received with the wrong Interface Identifier. That is, if a Hello is sent with a spoofed source address so that it appears to come from a known neighbor, and the Interface Identifier is different from what that neighbor is sending. Also, including this identifier in a spoofed message when the real neighbor is not sending it, or omitting it when the real neighbor is sending it. The effects of such attacks depend on how this Interface Identifier is used by other protocols.

5. IANA Considerations

IANA is requested to assign a PIM Hello Option value for the Interface Identifier option defined in this document.

6. Acknowledgments

The authors thank Yiqun Cai and Heidi Ou for providing valuable feedback.

7. References

7.1. Normative References

[I-D.ietf-pim-port]
Farinacci, D., Wijnands, I., Venaas, S., and M. Napierala, "A Reliable Transport Mechanism for PIM", draft-ietf-pim-port-04 (work in progress), October 2010.

[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.

7.2. Informative References

[HELLO-OPT]
IANA, "PIM Hello Options", PIM-HELLO-OPTIONS per RFC4601 http://www.iana.org/assignments/pim-hello-options, March 2007.

[RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets:MIB-II", STD 17, RFC 1213, March 1991.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

Authors' Addresses

Sameer Gulrajani
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: sameerg@cisco.com

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

Email: stig@cisco.com

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: September 6, 2011

Yiqun Cai
Liming Wei
Heidi Ou
Cisco Systems, Inc.

March 5, 2011

Protocol Independent Multicast ECMP Assert
draft-hou-pim-ecmp-00.txt

Abstract

A PIM router uses RPF procedure to select an upstream interface and router to build forwarding state. When there are equal cost multiple paths (ECMP), existing implementations often use hash algorithms to select a path. Such algorithms do not allow the spread of traffic among the ECMPs according to administrative metrics. This usually leads to inefficient or ineffective use of network resources. This document introduces the ECMP Assert, a mechanism to improve the RPF procedure over ECMPs. It allows ECMP path selection to be based on administratively selected metrics, such as data transmission delays, path preferences and routing metrics.

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 September 6, 2011.

Copyright Notice

Table of Contents

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].

2. Introduction

A PIM [RFC4601] router uses RPF procedure to select an upstream interface and a PIM neighbor on that interface to build forwarding state. When there are equal cost multiple paths (ECMP) upstream, existing implementations often use hash algorithms to select a path. Such algorithms do not allow the spread of traffic among the ECMP according to administrative metrics. This usually leads to inefficient or ineffective use of network resources. This document introduces the ECMP Assert, a mechanism to improve the RPF procedure over ECMP. It allows ECMP path selection to be based on administratively selected metrics, such as data transmission delays, path preferences and routing metrics, or a combination of metrics.

ECMPs are frequently used in networks to provide redundancy and to increase available bandwidth. A PIM router selects a path in the ECMP based on its own implementation specific choice. The selection is a local decision. One way is to choose the PIM neighbor with the highest IP address, another is to pick the PIM neighbor with the best hash value over the destination and source addresses.

While implementations supporting ECMP have been deployed widely, the existing RPF selection methods have weaknesses. The lack of administratively effective ways to allocate traffic over alternative paths is a major issue. For example, there is no straightforward way to tell two downstream routers to select either the same or different RPF neighbor routers for the same traffic flows.

With the ECMP Assert mechanism introduced here, the upstream routers use a new PIM ECMP Assert message to instruct the downstream routers on how to tie-break among the upstream neighbors. The PIM ECMP Assert message conveys the tie-break information based on metrics selected administratively.

2.1. Overview

The existing PIM Assert mechanism allows the upstream router to detect the existence of multiple forwarders for the same multicast flow onto the same downstream interface. The upstream router sends a PIM Assert message containing a routing metric for the downstream routers to use for tie-breaking among the multiple upstream

forwarders on the same RPF interface.

With ECMP interfaces between the downstream and upstream routers, the PIM ECMP Assert mechanism works in a similar way, but extends the ability to resolve the selection of forwarders among different interfaces in the ECMP.

When a PIM router downstream of the ECMP interfaces creates a new (*,G) or (S,G) entry, it will populate the RPF interface and RPF neighbor information according to the rules specified by [RFC4601]. This router will send its initial joins to that RPF neighbor.

When the RPF neighbor router receives the join message and finds that the receiving interface is one of the ECMP interfaces, it will check if the same flow is already being forwarded out of another ECMP interface. If so, this RPF neighbor router will send a PIM ECMP Assert message onto the interface the join was received on. The PIM ECMP Assert message contains the address of the desired RPF neighbor, an interface ID [INTID], along with other parameters used as tie breakers. In essence, a PIM ECMP Assert message is sent by an upstream router to notify downstream routers to redirect PIM Joins to the new RPF neighbor via a different interface. When the downstream routers receive this message, they should trigger PIM Joins toward the new RPF neighbor specified in the packet.

This new message is named PIM ECMP Assert for the following reasons,

1. It is sent by an upstream router;
2. It is used to influence the RPF selection by downstream routers; And
3. A tie breaker metric is used.

This new message functions in similar ways to the existing PIM Assert message, with the exception that the existing Assert message is used to select an upstream router within the same multi-access network (such as a LAN) while the new message is used to select both a network and an upstream router.

One advantage of this design is that the control messages are only sent when there is need to "re-balance" the traffic. This reduces the amount of control traffic.

2.2. Applicability

The use of ECMP Assert applies to shared trees or source trees built with procedures described in [RFC4601]. The use of ECMP Assert in "Protocol Independent Multicast - Dense Mode" [RFC3973] or in "Bidirectional Protocol Independent Multicast" [RFC5015] is not

considered.

The enhancement described in this document can be applicable to a number of scenarios. For example, it allows a network operator to use ECMP paths and have the ability to perform load splitting based on bandwidth. To do this, the downstream routers perform RPF selection with bandwidth instead of IP addresses as a tie breaker. The ECMP Assert mechanism assures that all downstream routers select the desired network link and upstream router whenever possible. Another example is for a network operator to impose a transmission delay limit on certain links. The ECMP Assert mechanism provides a mean for an upstream router to instruct a downstream router to choose a different RPF path.

This specification does not dictate the scope of applications of this mechanism.

3. Protocol Specification

3.1. ECMP Bundle

An ECMP bundle is a set of PIM enabled interfaces on a router, where all interfaces belonging to the same bundle share the same routing metric. The ECMP paths reside between the upstream and downstream routers over the ECMP bundle.

There can be one or more ECMP bundles on any router, while one individual interface can only belong to a single bundle.

ECMP bundles are created on a router via configuration.

3.2. Sending ECMP Assert

ECMP Asserts are sent by an upstream router in a rate limited fashion, under the following conditions,

- It detects a PIM Join on a non-desired outgoing interface; or
- It detects multicast traffic on a non-desired outgoing interface.

In both cases, an ECMP Assert is sent to the non-desired interface. An outgoing interface is considered "non-desired" when,

- The upstream router is already forwarding the same flow out of another interface belonging to the same ECMP bundle;
- The upstream router is not forwarding the flow yet out any interfaces of the ECMP bundle, but there is another interface with more desired attributes.

An upstream router may choose not to send ECMP Asserts if it becomes aware that some of the downstream routers do not support the new message.

3.3. Receiving ECMP Assert

When a downstream router receives an ECMP Assert, and detects the desired RPF path from its upstream router's point of view is different from its current one, it should choose to prune from the current path and join to the new path. The exact order of such actions is implementation specific.

If a downstream router receives multiple ECMP Asserts sent by different upstream routers, it SHOULD use the Preference, Metric, or other fields as specified below, as the tie breakers to choose the most preferred RPF interface and neighbor.

If an upstream router receives an ECMP Assert from another upstream router, it SHOULD NOT change its forwarding behavior even if the ECMP Assert makes it a less preferred RPF neighbor on the receiving interface.

3.4. Interoperability

If a PIM router supports this draft, it MUST send the new Hello option ECMP-Assert-Supported TLV in its PIM Hello messages. A PIM router sends ECMP Asserts on an interface only when it detects that all neighbors have sent this Hello option. If a PIM router detects that any of its neighbor does not support this Hello option, it MUST not send ECMP Asserts, however, it SHOULD still process any ECMP Asserts received.

3.5. Packet Format

3.5.1. PIM ECMP Assert Hello Option

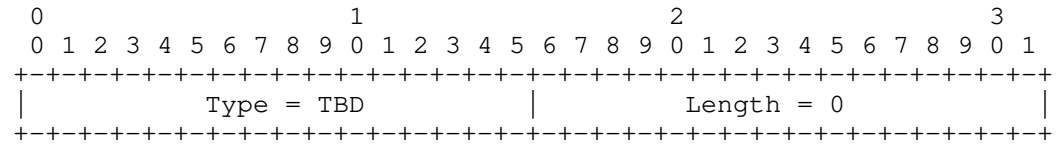


Figure 1: ECMP Assert Hello Option

Type: TBD.
Length: 0

3.5.2. PIM ECMP Assert Format

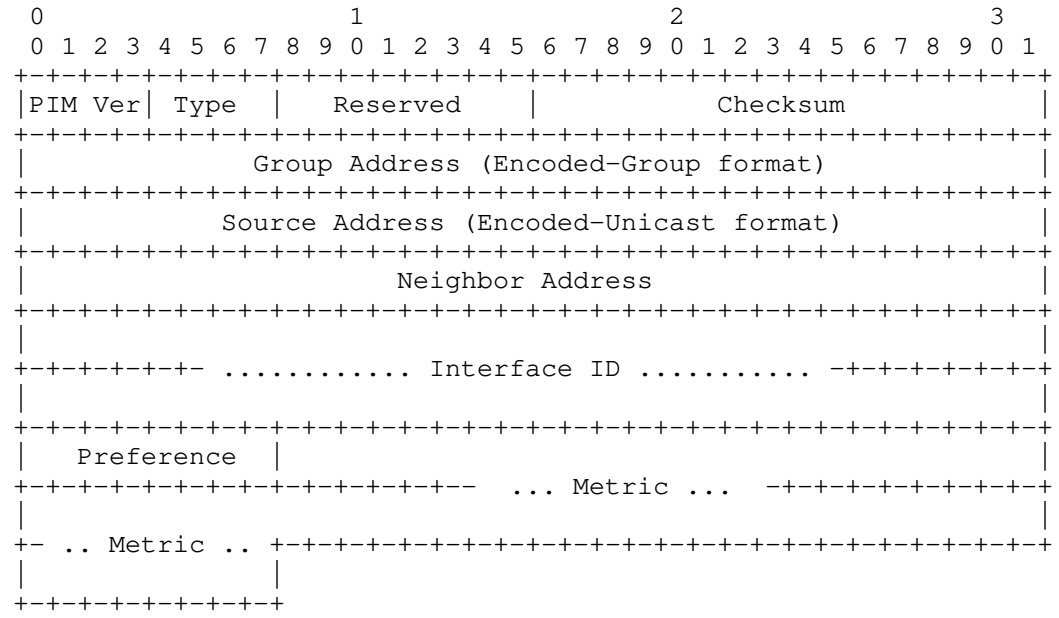


Figure 2: ECMP Assert Message Format

Type: TBD

Neighbor Address (32/128 bits): Address of desired upstream neighbor where the downstream receiver should redirect PIM Joins to. This address MUST be associated with an interface in the same ECMP bundle as the ECMP Assert message's outgoing interface. If the "Interface ID" field (see below) is ignored, this "Neighbor Address" field uniquely identifies a LAN and an upstream router to which a downstream router should redirect its Join messages to, and an ECMP Assert message MUST be discarded if the "Neighbor Address" field in the message does not match cached neighbor address.

Interface ID (64 bits): This field is used in IPv4 when one or more RPF neighbors in the ECMP bundle are unnumbered, or in IPv6 where link local addresses are in use. For other IPv4 usage, this field is zero'ed when sent, and ignored when received. If the "Router ID" part of the "Interface ID" is zero, the field must be ignored.

See [INTID] for details of its assignment and usage in PIM Hellos. If the "Interface ID" is not ignored, the receiving router of this message MUST use the "Interface ID", instead of "Neighbor Address", to identify the new RPF neighbor, and an ECMP Assert message MUST be discarded if the "Interface ID" field in the message does not match cached interface ID.

Preference (8 bits): The first tie breaker when ECMP Asserts from multiple upstream routers are compared against each other. Numerically smaller value is preferred. A reserved (15) value is used to indicate the metric value following the "Preference" field is a timestamp, taken at the moment the sending router started to forward out of this interface.

Metric (64 bits): The second tie breaker if the the "Preference" values are the same. Numerically smaller metric is preferred. This "Metric" can contain path parameters defined by users. When both "Preference" and "Metric" values are the same, "Neighbor Address" or "Interface ID" field is used as the third tie-breaker, depends on which field is used to identify the RPF neighbor, and the bigger value wins.

4. IANA Considerations

A new PIM Type is required to be assigned to the ECMP Assert messages. According to [PIMREG], this document recommends 11 (0xB) as the new "PIM ECMP Assert Type".

5. Security Considerations

Security of the ECMP Assert is only guaranteed by the security of the PIM packet, so the security considerations for PIM Assert packets as described in [RFC4601] apply here. Spoofed ECMP Assert packets may cause the downstream routers to send PIM Joins to an undesired upstream router, and trigger more ECMP Assert messages.

6. Acknowledgement

The authors would like to thank Apoorva Karan for helping with the original idea, Eric Rosen, Isidor Kouvelas, Toerless Eckert and Stig Venaas for their review comments.

7. References

7.1. Normative Reference

[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.

7.2. Informative References

[RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, January 2005.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

[INTID] Gulrajani, S. and S. Venaas, "PIM Interface ID Hello Option", draft-gulrajani-pim-hello-intid-00.txt (work in progress).

[PIMREG] Venaas, S., "A Registry for PIM Message Types", draft-ietf-pim-registry-04.txt (work in progress).

Authors' Addresses

Yiqun Cai
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: ycai@cisco.com

Liming Wei
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: lwei@cisco.com

Heidi Ou
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: hou@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 5, 2011

B. Fenner
AT&T Labs--Research
B. Haberman
Johns Hopkins University Applied Physics Lab
H. Holbrook
Arastra, Inc.
I. Kouvelas
S. Venaas
cisco Systems
March 4, 2011

Multicast Source Notification of Interest Protocol (MSNIP) draft-ietf-magma-msnip-06.txt

Abstract

This document discusses the Multicast Source Interest Notification Protocol (MSNIP). MSNIP is an extension to IGMPv3 and MLDv2 that provides membership notification services for sources of multicast traffic operating within the SSM destination address range.

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 September 5, 2011.

Copyright Notice

Table of Contents

1. Introduction

The Multicast Source Notification of Interest Protocol (MSNIP) is an extension to version 3 of the Internet Group Membership Protocol (IGMPv3 [RFC3376]) and version 2 of the Multicast Listener Discovery Protocol (MLDv2 [RFC3810]). MSNIP operates between multicast sources and their first-hop routers to provide information on the presence of receivers to the source systems. Using the services offered by MSNIP an application on an IP system wishing to source multicast data can register to be notified when receivers join and leave the session. This enables multicast sources to avoid the work of transmitting packets onto their first-hop link when there are no joined receivers.

A common scenario where MSNIP may be useful is one where there is a multicast server offering a large pool of potential flows that map onto separate multicast destination addresses but receivers exist only for a small subset of the flows. If the source were to continuously transmit data for all the flows that could potentially have receivers, significant resources would be wasted in the system itself as well as the first-hop link and first-hop router. Using a higher level control protocol to determine the existence of receivers for particular flows may not be practical as there may be many potential receivers in each active session.

Information on which multicast destination addresses have receivers for a particular sender is typically available to the multicast routing protocol on the first hop router for a source. MSNIP uses this information to notify the application in the sending system of when it should start or stop transmitting. This is achieved without any destination address specific state on the first-hop router for potential sources without receivers.

2. Routing Protocol Support

For reasons described in this section, MSNIP only supports transmission control for applications that use multicast destination addresses that are routed using Source Specific Multicast (SSM). See Appendix A for information on how MSNIP potentially can be extended to also work with Any-Source Multicast (ASM).

Many currently deployed multicast routing protocols require data from an active source to be propagated past the first-hop router before information on the existence of receivers becomes available on the first-hop. In addition, such protocols require that this activity is repeated periodically to maintain source liveness state on remote routers. All dense-mode protocols fall under this category as well as sparse-mode protocols that use shared trees for source discovery (such as PIM-SM [RFC4601]). In order to provide receiver interest notification for such protocols, the default mode of operation would require that the source IP system periodically transmits on all potential destination addresses and the first-hop routers prune the traffic back. Such a flood-and-prune behavior on the first-hop link significantly diminishes the benefits of managing source transmission.

In contrast, with source-specific sparse-mode protocols such as PIM-SSM [RFC4601]) availability of receiver membership information on the first-hop routers is independent of data transmission. Such protocols use an external mechanism for source discovery (like source-specific IGMPv3 membership reports) to build source-specific multicast trees.

Clearly these two classes of routing protocols require different handling for the problem MSNIP is trying to solve. In addition the second type covers the majority of applications that MSNIP is targeted at. MSNIP avoids the extra complication in supporting routing protocols that require a flood and prune behavior.

3. Service Interface for Requesting Membership Notification

Applications within an IP system that wish to source multicast packets and are interested in being notified on the existence of receivers must register with the IP layer of the system. MSNIP requires that within the IP system, there is (at least conceptually) a service interface that can be used to register with the IP layer for such notifications. Dual stack systems supporting both IPv4 and IPv6 need to provide separate service interfaces for each protocol.

A system's IPv4 or IPv6 service interface must support the following operation or any logical equivalent:

IPMulticastSourceRegister (socket, source-address, multicast-address)

IPMulticastSourceDeregister (socket, source-address, multicast-address)

In addition the application must provide the following interface for receiving notifications from the IP system:

IPMulticastSourceStart (socket, source-address, multicast-address)

IPMulticastSourceStop (socket, source-address, multicast-address)

where:

socket: is an implementation-specific parameter used to distinguish amongst different requesting entities (e.g., programs or processes) within the system; the socket parameter of BSD UNIX system calls is a specific example.

source-address: is the IP unicast source address that the application wishes to use in transmitting data to the specified multicast address. The specified address must be one of the IP addresses associated with the network interfaces of the IP system. Note that an interface in an IP system may be associated with more than one IP address. An implementation may allow a special "unspecified" value to be passed as the source-address parameter, in which case the request would apply to the "primary" IP address of the "primary" or "default" interface of the system (perhaps established by system configuration). If transmission to the same multicast address is desired using more than one source IP address, IPMulticastSourceRegister must be invoked separately for each desired source address.

multicast-address: is the IP multicast destination address to which the request pertains. If the application wishes to transmit data to more than one multicast addresses for a given source address, IPMulticastSourceRegister must be invoked separately for each desired multicast address.

3.1. Application Operation

Applications wishing to use MSNIP to control their multicast data transmission to destination G from source address S operate as follows.

Initially the application contacts the IP system to obtain the socket handle for use on all subsequent interactions. The application invokes IPMulticastSourceRegister for the desired S and G and then waits without transmitting any packets for the IP system to notify that receivers for the session exist.

If and when the IP system notifies the application that receivers exist using the IPMulticastSourceStart call, the application may start transmitting data. After the application has been notified to send, if all receivers for the session leave, the IP system will notify the application using the IPMulticastSourceStop call. At this point the application should stop transmitting data until it is notified again that receivers have joined through another IPMulticastSourceStart call.

When the application no longer wishes to transmit data it should invoke the IPMulticastSourceDeregister call to let the IP system know that it is no longer interested in notifications for this source and destination. The IPMulticastSourceDeregister call should be implicit in the teardown of the associated socket state.

4. MSNIP Managed Address Range Negotiation

With current multicast deployment in the Internet, different multicast routing protocols coexist and operate under separate parts of the multicast address space. Multicast routers are consistently configured with information that maps specific multicast address ranges to multicast routing protocols. Part of this configuration describes the subset of the address space that is used by source-specific multicast (SSM) [RFC5771]. As described in section 2, MSNIP only tries to control application transmission within the SSM address range.

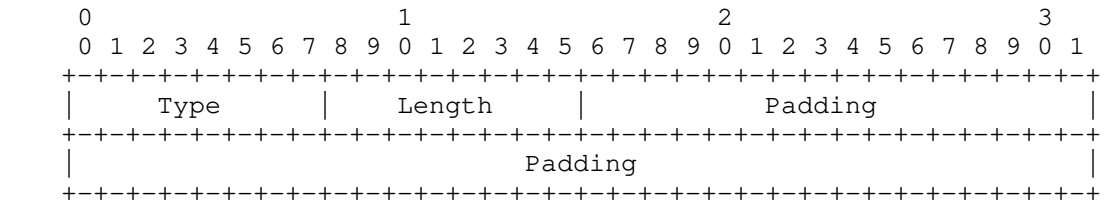
It is desirable for applications within an IP system that supports MSNIP to have a consistent service interface for multicast transmission that does not require the application to be aware of the SSM address range. MSNIP supports this by allowing applications to use the service interface described in section 3 for multicast destination addresses that are outside its operating range. When an application registers for notifications for a destination address that is not managed by MSNIP it is immediately notified to start transmitting. This is equivalent to the default behavior of IP multicast without MSNIP support which forces multicast applications to assume that there are multicast receivers present in the network.

4.1. Router Coordination

In order for MSNIP to operate on a shared link where two or more multicast routers may be present, all the multicast routers must be MSNIP-capable and have an identical configuration for the SSM address range. MSNIP enforces these requirements by using two new options for IPv4 in the Multicast Router Discovery protocol [RFC4286] and one new option for IPv6 in the Neighbor Discovery / ICMPv6 protocol [RFC4861].

4.1.1. MSNIP Operation Option

A multicast router advertises that it is participating in MSNIP using the MSNIP Operation option in either the Multicast Router Discovery protocol for IPv4 or the Neighbor Discovery / ICMPv6 protocol for IPv6. This option MUST be included in all router advertisement messages of a router that is configured for MSNIP. The format of the option is as follows:



Type: The type field is set to WW (TBD by IANA) for IPv4 and ZZ (TBD by IANA) for IPv6.

Length: The length field is set to 0 for IPv4 and 1 for IPv6.

Padding: The six extra bytes of padding are only present in IPv6 and are required to bring the size of the option up to the eight octet boundary. The value of the padding bytes must be set to zero on transmission and ignored on receipt.

A multicast router uses received Multicast Router Advertisement and Neighbor Discovery / ICMPv6 messages to determine if all live neighbor multicast routers on each interface are participating in MSNIP. When a router advertisement message not containing an MSNIP option is received by a router participating in MSNIP, the mis-configuration SHOULD be logged to the operator in a rate-limited manner.

If even one multicast router on a link does not have MSNIP capability then ALL routers on that link MUST be configured to not provide MSNIP services and to not advertise the MSNIP Operation option.

4.1.2. SSM Range Option

The SSM Range Multicast Router Discovery option advertises the IPv4 SSM Range with which the router is configured. The option is defined in [I-D.ietf-magma-mrdssm]. This option is only valid in IPv4. The SSM range for IPv6 is well defined for all valid scopes [RFC3306] and a mechanism to allow additional ranges to operate in SSM mode on a per-link bases is not required.

4.2. Managed Range Discovery by Source Systems

When an application in an IP system uses the MSNIP service interface to register for notification, the IP system must behave differently depending on whether or not the destination address for which the application registered is operating under SSM (and is being managed by MSNIP). For SSM channels, the IP system should only instruct the application to transmit when there are receivers for the multicast destination. For non-SSM destination addresses the IP system will

not be able to determine if there are receivers and should immediately instruct the application to transmit. In addition, an MSNIP-capable IP system must be able to detect if there are multicast routers on its connected links and if they support MSNIP operation. If no multicast routers are present or if the multicast routers are not MSNIP-capable then the IP system MUST default to flooding and immediately instruct applications to transmit.

An IP system controls transmission behavior under the different possible conditions by adapting its definition of the MSNIP-managed multicast destination address range:

- On a link with multicast routers operating the MSNIP protocol the IP system MUST use the SSM multicast destination address range as the MSNIP-managed range. IPv4 systems MUST use the contents of the SSM Range option in received Multicast Router Advertisement messages [I-D.ietf-magma-mrdssm] to discover the configured SSM range. SSM range discovery is not needed in IPv6 where the SSM destination address range is fixed.

- On a link not connected to a multicast routed infrastructure or on a link with multicast routers that do not support MSNIP operation, the IP system MUST use an empty range as its MSNIP-managed range. This forces applications transmitting to any multicast destination address to default to flooding thus providing backward compatibility.

As described in Section 4.1.1, an IP system can determine the status of a link and distinguish between the above two cases through the reception of IPv4 Multicast Router Advertisement and Neighbor Discovery / ICMPv6 messages.

5. Requesting and Receiving Notifications

Like IGMP, MSNIP is an asymmetric protocol specifying different behavior for systems wishing to source traffic and for multicast routers. Host IP systems multicast Host Interest Solicitation messages to register for notification with their first-hop routers. Routers unicast Router Receiver Membership Reports to IP systems to notify them of the arrival of the first or departure of the last receiver for a session. Note that a system may perform at the same time both of the above functions. An example is a router that wishes to source traffic.

5.1. Host Interest Solicitation

Source systems that wish to be managed by MSNIP periodically transmit a Host Interest Solicitation message. This message is multicast with a multicast destination address of ALL_IGMPv3_ROUTERS (224.0.0.22) or ALL_MLDv2_ROUTERS (FF02::16) and is transmitted every [Interest Solicitation Interval] seconds. The Host Interest Solicitation message contains a holdtime which is set to [Interest Solicitation Holdtime] and instructs the multicast first-hop routers to maintain MSNIP state for this IP system for the specified period. Systems with multiple interfaces or multiple IP addresses per interface must originate separate Host Interest Solicitation messages from each of their IP addresses that they wish to have managed by MSNIP. In practice a system with more than one IP address is treated by MSNIP as multiple IP systems.

When an IP system first comes up it transmits [Robustness Variable] Host Interest Solicitation messages spaced by [Initial Interest Solicitation Interval] seconds.

All MSNIP capable routers that receive a Host Interest Solicitation message from an IP system, maintain a system interest record of the form:

(IP system address, holdtime timer)

Each time a Host Interest Solicitation message is received from the IP system, the holdtime timer is reset to the holdtime in the received message. In addition the router may respond to the solicitation message with a Receiver Membership Report message described in Section 5.2. The message contains a TRANSMIT record for each of the multicast destination addresses within the MSNIP-managed range for which the routing protocol indicates there are receivers for this source system.

The holdtime timer of a record counts down to zero. When the

holdtime timer of a specific system interest record expires, the record is deleted.

5.2. Router Receiver Membership Reports

Receiver Membership Report messages are used by routers to communicate the receiver membership status of particular multicast destination addresses to a specific IP system. Each message contains a list of transmission control records of either TRANSMIT or HOLD type that instruct a system to respectively start or stop sending traffic on this link to the specified multicast destination address. Receiver Membership Report messages are unicast to the target system.

In addition to reports sent in response to Host Interest Solicitation messages, routers send unsolicited Receiver Membership Reports to IP systems when the receiver membership status reported by the multicast routing protocol changes for a specific source and multicast destination. Such reports are only sent if the multicast destination address is managed by MSNIP and the router has a system interest record created by a previously received Host Interest Solicitation message with an IP system address equal to the source address. If the source / destination pair satisfy these conditions then [Robustness Variable] Receiver Membership Reports are sent out spaced by [Unsolicited Membership Report Interval] seconds. If the membership status changes again for the same destination address and source system while transmission of Receiver Membership Reports is still pending then the pending report messages are canceled and a new set of [Robustness Variable] messages indicating the new state are scheduled.

When an IP system receives a Receiver Membership Report message, for each of the TRANSMIT records listed in the message, it creates or updates a transmission record of the form:

(router address, source address, multicast address, holdtime timer)

The router address is obtained from the source address of the IP header of the received message. The source address is obtained from the destination address of the IP header of the received message. The multicast address is obtained from the information in the TRANSMIT record. The holdtime timer is set to the value of the holdtime field in the received Receiver Membership Report message.

For each HOLD record present in the message, the system deletes the matching previously created transmission record from its state.

The holdtime timer of a record counts down to zero. When the

holdtime timer of a specific transmission record expires, the record is deleted.

Note that creation and deletion of transmission records in an IP system's state may cause local applications to be notified to start and stop transmitting data (see Section 6).

6. Application Notification

This section describes the relation between protocol events and notifications to source applications within an IP system. The state machine below is specific to each source address of the IP system and each multicast destination address. The initial state is the No Info state.

In tabular form, the state-machine is:

Event	Previous State		
	No Info	Hold	Transmit
New Register	- Start new	-	- Start new
Start Manage	-> Hold Stop ALL registered	-	-
Stop Manage	-	-> No Info Stop ALL registered	-> No Info
Recv TRANSMIT	-	-> Transmit Start ALL registered	-
Recv last HOLD or timeout	-	-	-> Hold Stop ALL registered

The events in the state machine above have the following meaning:

New register: A new application has registered through a call to IPMulticastSourceRegister for this S and G.

Start manage: We received an SSM Range option in an MRD packet on the interface that S belongs to that changed the status of G from a non-managed to a MSNIP-managed destination address. The SSM Range option is only valid in IPv4.

Stop manage: We received an SSM Range option in an MRD packet on the interface that S belongs to that changed the status of G from a MSNIP-managed to a non-managed destination address or the mapping state that caused this destination address to be managed expired. The SSM Range option is only valid in IPv4.

Receive TRANSMIT: We received a Receiver Membership Report with S as the IP destination address that contains a TRANSMIT record for G.

Receive last HOLD or timeout: We either received a Receiver Membership Report with S as the IP destination address that contains a HOLD record for G or the holdtime timer in a transmission record for S and G expired and there are no other transmission records for S and G. This means that the last router that was reporting receivers no longer does so and there are no routers left wishing to receive traffic from this S to destination address G.

The state machine actions have the following meaning:

Start new: Send an IPMulticastSourceStart notification to the application that just registered for this S and G.

Start ALL registered: Send an IPMulticastSourceStart notification to all applications that are registered for this S and G.

Stop ALL registered: Send an IPMulticastSourceStop notification to all applications that are registered for this S and G.

7. Router Processing

This section describes the per-source system tracking state machine operated by each first-hop router. The initial state is No Info.

In tabular form, the state-machine is:

Event	Previous State	
	Not tracking	Tracking
Receive HIS	-> Tracking Set HT to message holdtime; Send ALL existing TRANSMITs	- Set HT to message holdtime; Send ALL existing TRANSMITs
HIS timeout	-	-> Not tracking
Receivers for new destination G	-	- Send TRANSMIT for G
Receivers of G leave	-	- Send HOLD for G

The events in the state machine above have the following meaning:

Receive HIS: The router has received a Host Interest Solicitation from S.

HIS timeout: The holdtime timer (HT) in the host interest record associated with S has expired.

Receivers for new destination G: The routing protocol has informed MSNIP that it now has receivers for the MSNIP-managed destination address G and source IP system S.

Receivers of G leave: The routing protocol has informed MSNIP that all receivers for the MSNIP-managed destination address G and source IP system S have left the channel.

The state machine actions have the following meaning:

Set HT to message holdtime: The holdtime timer in the host interest record associated with S is restarted to the value of the holdtime field in the received Host Interest Solicitation message.

Send ALL existing TRANSMITs: The router builds and transmits Receiver Membership Reports to S that contain a TRANSMIT record for each of the MSNIP-managed destination addresses that have receivers for S.

Send TRANSMIT for G: The router builds and transmits a Receiver Membership Report to S that contains a TRANSMIT record for the destination address G.

Send HOLD for G: The router builds and transmits a Receiver Membership Report to S that contains a HOLD record for the destination address G.

8. Message Formats

The following packet formats are valid for both IPv4 and IPv6. IP version-specific values will be explicitly defined.

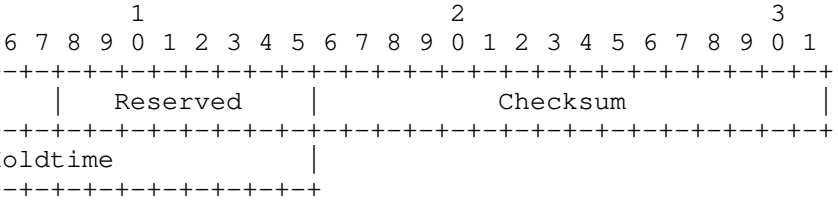
There are two message types of concern to the MSNIP protocol described in this document:

Type Number (hex)	Message Name
0xXX	Host Interest Solicitation
0xYY	Receiver Membership Report

Both the Host Interest Solicitation message and the Receiver Membership Report message MUST not be forwarded by routers (see Section 12). The Router Alert option [RFC2113] [RFC2711] MUST be included in the packet by the router or host IP system transmitting the message. Routers receiving Host Interest Solicitation messages and IP systems receiving Receiver Membership Reports MUST not process a received MSNIP message if the Router Alert option is not present.

8.1. Host Interest Solicitation Message

A Host Interest Solicitation message is periodically multicast by MSNIP capable systems to declare interest in Receiver Membership Reports from multicast routers on the link. The Host Interest Solicitation message is multicast with a destination address of ALL_IGMPv3_ROUTERS (224.0.0.22) or ALL_MLDv2_ROUTERS (FF02::16).



Type: The type field is set to XX (to be assigned by IANA as an IGMP type for IPv4 and an ICMPv6 type for IPv6).

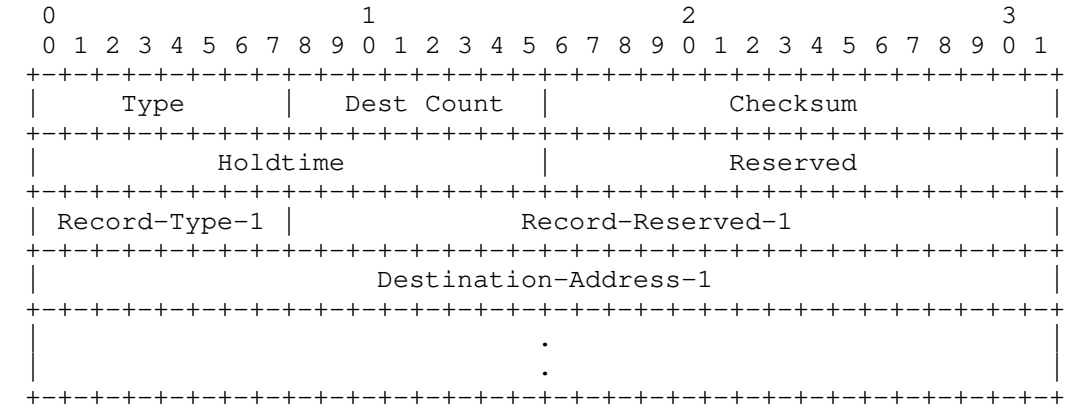
Reserved: Transmitted as zero. Ignored upon receipt.

Checksum: In IPv4, the Checksum is the 16-bit one's complement of the one's complement sum of the whole IGMP message (the entire IP payload). In IPv6, the Checksum is the standard ICMPv6 checksum, covering the entire MLDv2 message plus a "pseudo-header" of IPv6 header fields [RFC4443]. For computing the checksum, the Checksum field is set to zero. When receiving packets, the checksum MUST be verified before processing a packet.

Holdtime: The amount of time a receiving router must keep the system interest state alive, in seconds. The default value for this field is [Interest Solicitation Holdtime].

8.2. Receiver Membership Report Message

A Receiver Membership Report message is unicast by first-hop multicast routers and targeted at potential sources to inform them of the existence or not of receivers for the listed multicast destination addresses.



Type: The type field is set to YY (to be assigned by IANA as an IGMP type for IPv4 and an ICMPv6 type for IPv6).

Dest Count: The number of multicast destination address records present in this message.

Checksum: In IPv4, the Checksum is the 16-bit one's complement of the one's complement sum of the whole IGMP message (the entire IP payload). In IPv6, the Checksum is the standard ICMPv6 checksum, covering the entire MLDv2 message plus a "pseudo-header" of IPv6 header fields [RFC4443]. For computing the checksum, the Checksum field is set to zero. When receiving packets, the checksum MUST be verified before processing a packet.

Holdtime: The amount of time in seconds that the target host must keep alive the transmission record state created or updated by the TRANSMIT records in this report. The router originating the Receiver Membership Report sets this field to the current value of the holdtime timer in the system interest record corresponding to the target host. As a result Receiver Membership Reports sent in response to the reception of a Host Interest Solicitation message have their holdtime set to the value of the holdtime field in the received HIS message.

Reserved: Transmitted as zero. Ignored upon receipt.

Record-Type-1: The type of the first transmission control record in this message. Valid values are:

Record Type	Description	Value
TRANSMIT	Request to start transmitting to destination	1
HOLD	Request to stop transmitting to destination	2

Record-Reserved-1: Transmitted as zero. Ignored upon receipt.

Destination-Address-1: The multicast destination address of the first record in the message.

8.3. IPv4 Header Fields

Like all IGMP messages, MSNIP messages are encapsulated in IPv4 datagrams, with an IP protocol number of 2. MSNIP messages can be identified from other IGMP messages by the message types listed in Section 8. Every MSNIP message described in this document is sent with an IP Time-to-Live of 1, and carries an IP Router Alert option [RFC2113] in its IP header.

8.4. IPv6 Header Fields

MLD messages are a sub-protocol of the Internet Control Message Protocol (ICMPv6 [RFC4443]). MSNIP messages are identified in IPv6 packets by the combination of a preceding Next Header value of 58 and by the MLD message types listed in Section 8. All MSNIP messages described in this document are sent with a link-local IPv6 Source Address (or the unspecified address, if a valid link-local address is not available), an IPv6 Hop Limit of 1, and an IPv6 Router Alert option [RFC2711] in a Hop-by-hop Options header.

9. Constants Timers and Default Values

Robustness Variable: The Robustness Variable allows tuning for the expected packet loss on a network. If a network is expected to be lossy, the Robustness Variable may be increased. MSNIP is robust to (Robustness Variable - 1) packet losses. The Robustness Variable MUST NOT be zero, and SHOULD NOT be one. Default: 2

Interest Solicitation Interval: The interval used by MSNIP capable systems between transmissions of Host Interest Solicitation messages. Default: 60 secs

Interest Solicitation Holdtime: The interval inserted in Host Interest Solicitation messages by systems to instruct routers how long they should maintain system interest state for. This MUST be ((the Robustness Variable) times (the Interest Solicitation Interval) plus (one second)).

Initial Interest Solicitation Interval: The interval used by systems to send out the initial Host Interest Solicitation messages when they first come up. Default: 1 second.

Unsolicited Membership Report Interval: The interval used by routers to send out a set of Membership Report messages when the receiver membership changes for a specific system. Default: 1 second.

10. Possible Optimisations

10.1. Suppressing HIS Messages

A possible optimisation for MSNIP is to suppress the transmission of Host Interest Solicitation messages from the source address of an IP system for which no local application has registered interest. In addition to conserving bandwidth, not transmitting HIS messages prevents remote receivers for groups with no matching source application from creating transmission record state in the host system.

10.2. Host Stack Filtering

Legacy applications that have not been coded with MSNIP support can still be prevented from wasting first-hop link bandwidth by filtering transmitted packets at the operating system level. Even though such applications will not register for MSNIP notifications with the host operating system, if the OS is MSNIP-capable and the application is transmitting data to an MSNIP-managed group for which there are no transmit records, the OS can safely filter the packets and not transmit them on the wire.

A problem with the filtering approach is that it cannot be combined with the HIS message suppression optimisation (see Section 10.1). If there are no registered applications in the system and HIS messages are being suppressed then the first-hop routers will not send any Receiver Membership Reports to the system. As a result, knowledge of receiver membership from the presence of transmit records for groups operated by legacy applications will not exist. It therefore becomes unsafe to filter packets from legacy applications.

10.3. Responding to Unexpected IGMP Queries

Under steady state the router side of the IGMP protocol elects a single router on each link that is responsible for issuing IGMP Queries. Routers other than the acting IGMP querier will send an IGMP Query only if they restart and have no IGMP querier election state or if the active Querier crashes and a new election takes place.

MSNIP can take advantage of this mechanism to quickly populate the host interest records of a new router starting up. When the router comes up it will issue an IGMP Query in an attempt to be elected as a Querier. MSNIP-capable hosts will notice that the sender of the Query is not the acting Querier. They can use this trigger to respond with Host Interest Solicitation Messages (with transmission randomised over a small interval) to quickly bring the new router up-

to-date.

10.4. Host and Router Startup

When a host operating system is restarted there may be applications that are started as part of the initialisation process and want to source IPv4 multicast traffic. It is possible for the applications to register through MSNIP with the IP subsystem and to start transmitting multicast data before the host receives the MSNIP-managed range definition through the SSM Range option of the Multicast Router Discovery protocol.

This temporary flooding can be avoided if the host OS holds off notifying MSNIP-capable applications that they can transmit until it receives an MRD advertisement and learns the SSM configuration for the network. This behaviour has the drawback that it is not compatible with legacy networks with no MRD deployment. In such a network the host OS has to be able to determine after a configurable period that MRD is not enabled and hence all multicast applications wishing to source traffic should be notified to transmit. A good default value for this period is the MAX_RESPONSE_DELAY of the Multicast Router Discovery protocol [RFC4286].

Late router startup is harder to deal with. Hosts that start up before the multicast router may time out waiting for an MRD advertisement and instruct all MSNIP-capable multicast source applications to transmit data. One way to work around this problem is to configure the host OS to wait forever for an MRD advertisement before instructing MSNIP applications to transmit.

11. Inter-operation with IGMP / MLD Proxying

MSNIP is intended for use on networks with multicast servers offering a large number of potential sessions. Although unlikely, it is possible to deploy such a server behind an IGMP / MLD Proxy [RFC4605]. If the proxy is not MSNIP-aware and does not implement the extensions described below then sources behind the proxy will default to flooding.

If a device performing IGMP / MLD Proxying wishes to proxy MSNIP, it MUST forward MSNIP Host Interest Solicitation messages that are received on downstream interfaces to its upstream interface. No special treatment is required for MSNIP Receiver Membership Reports as they are unicast to the target host.

In addition to the forwarding of MSNIP messages, an IGMP proxy MUST operate the Multicast Router Discovery protocol [RFC4286] on all its downstream interfaces and advertise the MSNIP capability option (Section 4.1.1) and SSM address range option (Section 4.1.2). The MSNIP capability option should be advertised on downstream interfaces only if it is included in MRD messages received on the upstream interface. The address range to be included in the SSM Range option MUST be determined by MRD and IGMP messages received on the upstream interface of the proxy according to the rules in Section 4.2.

In addition to the forwarding of MSNIP messages, an MLD proxy MUST operate the IPv6 Neighbour Discovery protocol. The MSNIP capability option should be advertised on downstream interfaces when it is included in IPv6 Neighbour Discovery messages received on the upstream interface.

12. Security Considerations

We consider the ramifications of a forged message of each type. As described in [RFC3376] and [RFC3810], IPSEC AH can be used to authenticate IGMP and MLD messages if desired.

12.1. Receiver Membership Report Attacks

A DoS attack on a host could be staged through forged Receiver Membership Report messages. The attacker can send a large number of reports, each with a large number of TRANSMIT records and a holdtime field set to a large value. The host will have to store and maintain the transmission records specified in all of those reports for the duration of the holdtime. This would consume both memory and CPU cycles in the host.

Forged Receiver Membership Report messages from the local network can be easily traced. There are three measures necessary to defend against externally forged reports:

- Routers SHOULD NOT forward Receiver Membership Reports. This is easier for a router to accomplish if the report carries the Router-Alert option.

- Hosts SHOULD ignore Receiver Membership Reports without the Router-Alert option.

Note that a remote attack through the multicast routing protocol is possible. A remote site can originate join state for a large number of groups that will propagate through MSNIP to the target source host. Such attacks are considered a more significant problem for the routers involved and are left up to the routing protocol security.

HOLD records in forged Receiver Membership Report messages are not a significant threat as hosts track the individual interests of each first-hop router separately. Only by forging the source address of the report message so that is appears to have originated from a real first-hop router can the attacker cause the source to stop transmitting to a group that has valid receivers. Such forged messages can be detected by the router itself.

12.2. Host Interest Solicitation Attacks

Forged Host Interest Solicitation messages can have two effects:

- When non-existent source addresses are used the solicitation messages can create unwanted host record state on attached routers for the duration of the holdtime specified in the message.

- When a source address corresponding to an existing host is used in the forged HIS message, receipt of the message by attached routers will cause them to transmit Receiver Membership Reports messages for all MSNIP-managed multicast destination addresses with receivers for the target host. Although no additional state will be created in routers or hosts from this attack, bandwidth and CPU is wasted in both the first-hop routers and the target host.

Just like for the Receiver Membership Report message, attacks using the Host Interest Solicitation message can be reduced by requiring the use of the Router-Alert option on the message.

12.3. MSNIP Managed Range Discovery

As discussed in [I-D.ietf-magma-mrdssm] it is possible for directly connected systems to send forged Multicast Router Advertisement messages containing the SSM Range Discovery option. As the SSM Range Discovery option determines the MSNIP-managed range under IPv4, such forged messages can temporarily replace the managed range map with incorrect information in receiving hosts. An incorrect mapping can have two effects:

- Applications using a multicast destination address within the real SSM range that have no valid receivers can be tricked into thinking that their chosen destination address is no longer an SSM address and will therefore start transmitting data.

- Applications using group addresses outside the valid SSM range can be tricked into thinking that they are using an SSM destination address and therefore prevented from transmitting data.

The Multicast Router Discovery SSM Range Option specification suggests that a router receiving a Multicast Router Advertisement with an inconsistent SSM Range Option log the event to the operator. Such logging will enable tracking of this type of attack.

13. IANA Considerations

This document introduces the following new types and options that require allocation by IANA:

- Two new IGMP messages for Host Interest Solicitation and Receiver Membership Report. Each of these messages requires a new IGMP type value to be assigned by IANA [IGMPREG].

- The new MSNIP Operation option for the Multicast Router Discovery protocol. This option requires a new MRD type value to be assigned by IANA.

- The new MSNIP Operation option for the Neighbour Discovery / ICMPv6 protocol. This option requires a new NDP / ICMPv6 type value to be assigned by IANA.

14. Acknowledgements

The authors would like to thank Dave Thaler, Jon Crowcroft, Toerless Eckert, Haixiang He, Pekka Savola and Karen Kimball for their contribution to this specification.

15. References

15.1. Normative References

[I-D.ietf-magma-mrdssm]
Kouvelas, I., "Multicast Router Discovery SSM Range Option", draft-ietf-magma-mrdssm-03 (work in progress), June 2003.

[RFC2113] Katz, D., "IP Router Alert Option", RFC 2113, February 1997.

[RFC2711] Partridge, C. and A. Jackson, "IPv6 Router Alert Option", RFC 2711, October 1999.

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.

[RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.

[RFC4286] Haberman, B. and J. Martin, "Multicast Router Discovery", RFC 4286, December 2005.

[RFC4443] Conta, A., Deering, S., and M. Gupta, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 4443, March 2006.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

15.2. Informative References

[IGMPREG] IANA, "IGMP Type Numbers", IGMP TYPE NUMBERS - per RFC3228,
BCP57 http://www.iana.org/assignments/igmp-type-numbers, June 2005.

[RFC3306] Haberman, B. and D. Thaler, "Unicast-Prefix-based IPv6 Multicast Addresses", RFC 3306, August 2002.

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

[RFC4605] Fenner, B., He, H., Haberman, B., and H. Sandick,

"Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.

[RFC5771] Cotton, M., Vegoda, L., and D. Meyer, "IANA Guidelines for IPv4 Multicast Address Assignments", BCP 51, RFC 5771, March 2010.

Appendix A. Extending MSNIP to Any-Source Multicast

This document defines MSNIP only for use with SSM. As noted in Section 2 many currently deployed multicast routing protocols require data from an active source to be propagated past the first-hop router before information on the existence of receivers becomes available on the first-hop. We will specify in Appendix A.1 how MSNIP can be extended to work for ASM when PIM-SM [RFC4601]) is used.

Whether MSNIP can be used for ASM depends on the multicast routing protocols used. There may be different protocols used for different group addresses. Rather than requiring a host to know for which ASM groups MSNIP can be used, we suggest that the host can use it for all ASM groups. If the first-hop router is unable to determine whether there are receivers or not, it can tell the source that there are receivers present anyway. The host will then start sending and the behavior will be as if MSNIP is not used. If MSNIP is extended to ASM, one should consider adding a flag to the MSNIP Operation Option Section 4.1.1, or creating a new option for use with IPv4 in the Multicast Router Discovery protocol [RFC4286] and Neighbor Discovery / ICMPv6 protocol [RFC4861], in order to announced the router capability to the hosts.

A.1. Extending MSNIP to ASM with PIM-SM

When PIM-SM [RFC4601] is used to provide ASM service, a first-hop router will generally not know if there are receivers for a group until it starts receiving data from an active source. Until the source becomes active, receivers simply join the shared tree for the group. This allows the Rendezvous-Point (RP) for the group to learn that receivers are present. Next when a source becomes active, a first-hop router (the Designated Router (DR)) will be responsible for sending PIM register messages to the the RP. If there are receivers present, the RP and/or last-hop routers will join the Shortest Path Tree (SPT) towards the source. This will result in at least one first-hop router learning that a source exists. The last part is similar to when using PIM-SM for SSM. With SSM a last-hop router immediately joins the Shortest Path Tree (SPT).

MSNIP can be extended to ASM with PIM-SM as follows:

- Host Interest Solicitation Message (Section 8.1) need to be extended to include a list of groups that the host is interested in receiving membership reports for.

- When a Designated Router (DR) receives a Host Interest Solicitiation Message with source address S containing a group G, it will periodically send PIM Null-Register messages to the RP for

(S,G). This is done instead of the data PIM Register messages the DR would use if the source did not use MSNIP. Per the DR register state machine in section 4.4.1 of [RFC4601], one can immediately send a Null-Register and then move to Prune state as if a Register-Stop was received. When the Register-Stop timer expires, send a Null-Register as usual. But then, rather than setting the Register-Stop timer to Register_Probe_Time, transition directly to Prune state as if a Register-Stop was received again. By periodically receiving the (S,G) registers, the RP will know that a source exists, and will join the SPT towards the source if it has receivers. Just like SSM, a first-hop routers will then receive an SPT join for (S,G) and learn that there are receivers. It can then inform the source. If the first-hop router has (*,G)-state, e.g., local interest or it is part of the shared tree, but has not yet got an (S,G) olist, it must immediately inform the source.

One benefit with this approach, is that PIM data registers can be avoided.

Authors' Addresses

Bill Fenner
AT&T Labs--Research
1 River Oaks Place
San Jose, CA 95134
USA

Email: fenner@research.att.com

Brian Haberman
Johns Hopkins University Applied Physics Lab
11100 Johns Hopkins Road
Laurel, MD 20723-6099
USA

Email: brian@innovationslab.net

Hugh Holbrook
Arastra, Inc.
P.O. Box 10905
Palo Alto, CA 94303
USA

Email: holbrook@arastra.com

Isidor Kouvelas
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: kouvelas@cisco.com

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

Email: stig@cisco.com

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: September 12, 2011

Dino Farinacci
Greg Shepherd
Yiqun Cai
Stig Venaas
cisco Systems

March 11, 2011

Population Count Extensions to PIM
draft-ietf-pim-pop-count-03.txt

Abstract

This specification defines a method for providing multicast distribution-tree accounting data. Simple extensions to the PIM protocol allow a rough approximation of tree-based data in a scalable fashion.

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 September 12, 2011.

Copyright Notice

Table of Contents

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].

2. Introduction

This draft proposes a mechanism to convey accounting information using the PIM protocol [RFC4601] [RFC5015]. Putting the mechanism in PIM allows efficient distribution and maintenance of such accounting information. Previous mechanisms require data to be correlated from multiple router sources.

This proposal allows a single router to be queried to obtain accounting and statistic information for a multicast distribution tree as a whole or any distribution sub-tree downstream from a queried router. The amount of information is fixed and does not increase as multicast membership, tree diameter, or branching increase.

The sort of accounting data this draft provides, on a per multicast route basis, are:

1. The number of branches in a distribution tree.

2. The membership type of the distribution tree, that is SSM or ASM.

3. Routing domain and time zone boundary information.

4. On-tree node and tree diameter counters.

5. Effective MTU and bandwidth.

This draft adds a new PIM Join Attribute type [RFC5384] to the Join/ Prune message as well as a new Hello TLV. The mechanism is applicable to IPv4 and IPv6 multicast.

2.1. Terminology

This section defines the terms used in this draft.

Multicast Route: A (S,G) or (*,G) entry regardless if the route is in ASM, SSM, or Bidir mode of operation.

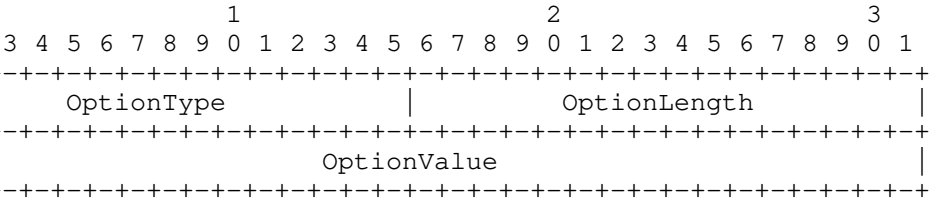
Stub Link: A link with members joined to the group via IGMP or MLD.

Transit Link: A link put in the oif-list for a multicast route because it was joined by PIM routers.

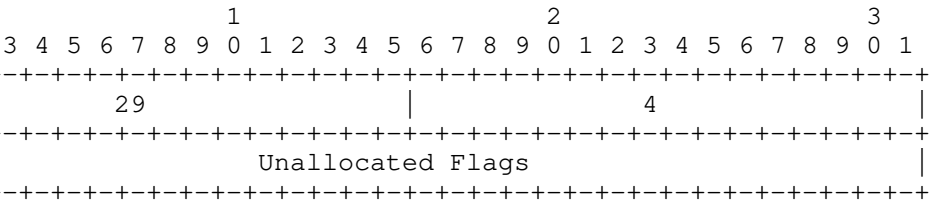
Note that a link can be both a Stub Link and a Transit Link at the same time.

3. New Hello TLV Pop-Count Support

When a PIM router sends a Join/Prune message to a neighbor, it will encode the data in a new PIM Join Attribute type (described in this draft) when the PIM router determines the neighbor can support this draft. If a PIM router supports this draft, it must send the Pop-Count-Supported TLV. The format of the TLV is defined to be:



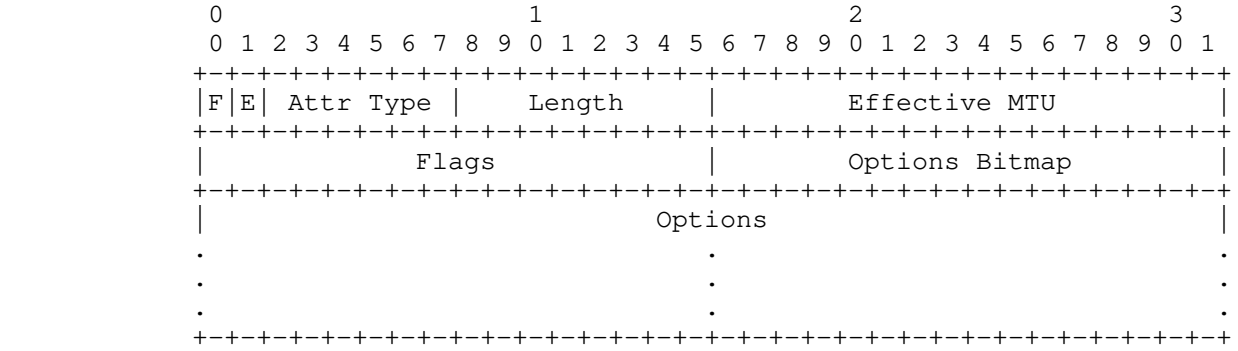
OptionType = 29, OptionLength = 4, there is no OptionValue semantics defined at this time but will be included for expandability and be defined in future revisions of this draft. The format will look like:



Unallocated Flags: for now should be sent as 0 and ignored on receipt.

4. New Pop-Count Join Attribute Format

When a PIM router supports this draft and has determined from a received Hello, the neighbor supports this draft, it will send Join/Prune messages that MAY include a Pop-Count attribute. The mechanism to process PIM Join Attribute is described in [RFC5384]. The format of the new attribute is described in the following.



The above format is used only for entries in the join-list section of the Join/Prune message.

F bit: 0 Non-Transitive Attribute.

E bit: As specified by [RFC5384].

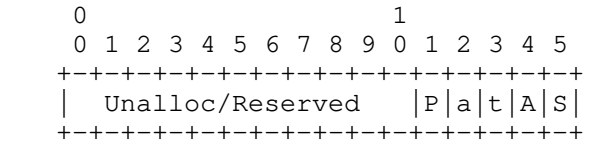
Attr Type: 2.

Length: The minimum length is 6.

Effective MTU: This contains the minimum MTU for any link in the oif-list. The sender of Join/Prune message takes the minimum value for the MTU (in bytes) from each link in the oif-list. If this value is less than the value stored for the multicast route (the one received from downstream joiners) then the value should be reset and sent in Join/Prune message. Otherwise, the value should remain unchanged.

This provides one to obtain the MTU supported by multicast distribution tree when examined at the first-hop router(s) or for sub-tree for any router on the distribution tree.

Flags: The flags field has the following format:



Unallocated Flags: The flags which are currently not defined. If a new flag is defined and sent by a new implementation, an old implementation should preserve the bit settings. This means that if a bit was set in a PIM Join message from any of the downstream routers, then it MUST also be set in any PIM Join sent upstream.

S flag: If an IGMPv3 or MLDv2 report was received on any oif-list entry or the bit was set from any PIM Join message. This bit should only be cleared when the above becomes untrue.

A flag: If an IGMPv1, IGMPv2, or MLDv1 report was received on any oif-list entry or the bit was set from any PIM Join message. This bit should only be cleared when the above becomes untrue.

A combination of settings for these bits indicate:

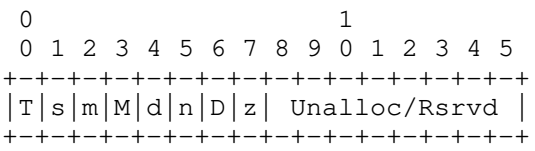
A-flag	S-flag	Description
0	0	There are no members for the group ('Stub Oif-List Count' is 0)
0	1	All group members are only SSM capable
1	0	All group members are only ASM capable
1	1	There is a mixture of SSM and ASM capable

t flag: If there are any tunnels on the distribution tree. If a tunnel is in the oif-list, a router should set this bit in its Join/Prune messages. Otherwise, it propagates the bit setting from downstream joiners.

a flag: If there are any auto-tunnels on the distribution tree. If an auto-tunnel is in the oif-list, a router should set this bit in its Join/Prune messages. Otherwise, it propagates the bit setting from downstream joiners. An example of an auto-tunnel is an tunnel setup by the AMT [AMT] protocol.

P flag: This flag remains set if all downstream routers support this specification. That is, they are PIM pop-count capable. This allows one to tell if the entire sub-tree is completely accounting capable.

Options Bitmap: This is a bitmap that shows which options are present. The format of the bitmap is as follows:



Each one of the bits T, s, m, M, d, n, D and z is associated with one option, where the option is included if and only if the respective bit is set. Included options MUST be in the same order as these bits are listed. The bits denote the following options:

bit	Option
T	Transit Oif-List Count
s	Stub Oif-List Count
m	Minimum Speed Link
M	Maximum Speed Link
d	Domain Count
n	Node Count
D	Diameter Count
z	TZ Count

See Section 4.1 for details on the different options. The unallocated bits are reserved. Any unknown bits MUST be set to 0 when a message is sent, and treated as 0 (ignored) when received. This means that unknown options which are denoted by unknown bits are ignored.

Options: This field contains options. Which options are present are determined by the flag bits. As new flags and options may be defined in the future, any unknown/reserved flags MUST be ignored, and any additional trailing options MUST be ignored. See Section 4.1 for details on the options defined in this document.

4.1. Options

There are several options defined in this document. For each option, there is also a related flag that shows whether the option is present. See the Options Bitmap above for a list of the options and their respective bits. Each option has a fixed size.

Transit Oif-List Count: This is filled in by a router sending a Join/Prune message which is equal to the number of oifs for the multicast route that has been joined by PIM. This indicates the transit branches on a multicast distribution tree (no members on the links between this router and joining routers). This is added to the value advertised by all downstream PIM routers that have joined on this oif. Length 2 octets.

Stub Oif-List Count: This is filled in by a router sending a Join/Prune message which is equal to the number of oifs for the multicast route that has been joined by IGMP or MLD. This indicates the links where there are host members for the multicast route. This is added to the value advertised by all downstream PIM routers that have joined on this oif. Length 2 octets.

Minimum Speed Link: This contains the minimum bandwidth rate for any link in the oif-list and is encoded as specified in Section 4.1.1. The sender of Join/Prune message takes the minimum value for each link in the oif-list for the multicast route. If this value is less than the value stored for the multicast route (the one received from downstream joiners) then the value should be reset and sent in Join/Prune message. Otherwise, the value should remain unchanged. This together with the Maximum Speed Link option provides a way to obtain the lowest and highest speed link for the multicast distribution tree. Length 2 octets.

Maximum Speed Link: This contains the maximum bandwidth rate for any link in the oif-list and is encoded as specified in Section 4.1.1. The sender of Join/Prune message takes the maximum value for each link in the oif-list for the multicast route. If this value is greater than the value stored for the multicast route (the one received from downstream joiners) then the value should be reset and sent in Join/Prune message. Otherwise, the value should remain unchanged. This together with the Minimum Speed Link option provides a way to obtain the lowest and highest speed link for the multicast distribution tree. Length 2 octets.

Domain Count: This indicates the number of routing domains the distribution tree traverses. A router should increment this value if it is sending a Join/Prune message over a link which traverses a domain boundary. Length 1 octet.

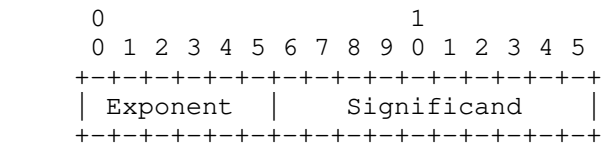
Node Count: This indicates the number of routers on the distribution tree. Each router will sum up all the Node Counts from all joiners on all oifs and increment by 1 before including this value in the Join/Prune message. Length 1 octet.

Diameter Count: This indicates the longest length of any given branch of the tree in router hops. Each router that sends a Join increments the max value received by all downstream joiners by 1. Length 1 octet.

TZ Count: This indicates the number of timezones the distribution tree traverses. A router should increment this value if it is sending a Join/Prune message over a link which traverses a time zone. This can be a configured link attribute or use other means to determine the timezone is acceptable. Length 1 octet.

4.1.1. Link Speed Encoding

The speed is encoded using 2 octets as follows:



Using this format, the speed of the link is Significand * 10 ^ Exponent kbps. This allows specifying link speeds with up to 3 decimal digits precision and speeds from 1 kbps to 10 ^ 67 kbps. A computed speed of 0 kbps means the link speed is < 1 kbps.

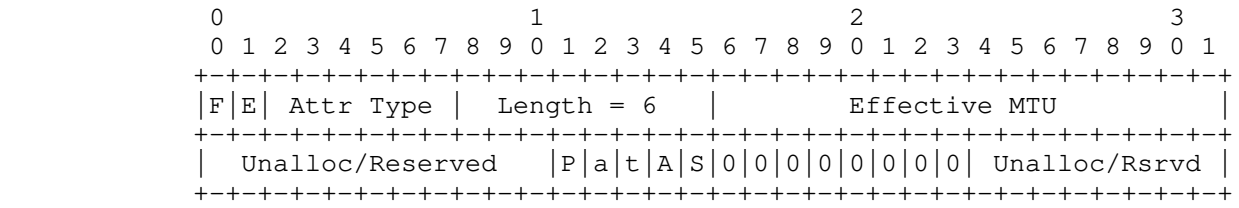
Here are some examples how this is used:

Link Speed	Exponent	Significand
500 kbps	0	500
500 kbps	2	5
155 Mbps	3	155
40 Gpbs	6	40
100 Gpbs	6	100
100 Gpbs	8	1

4.2. Example message layouts

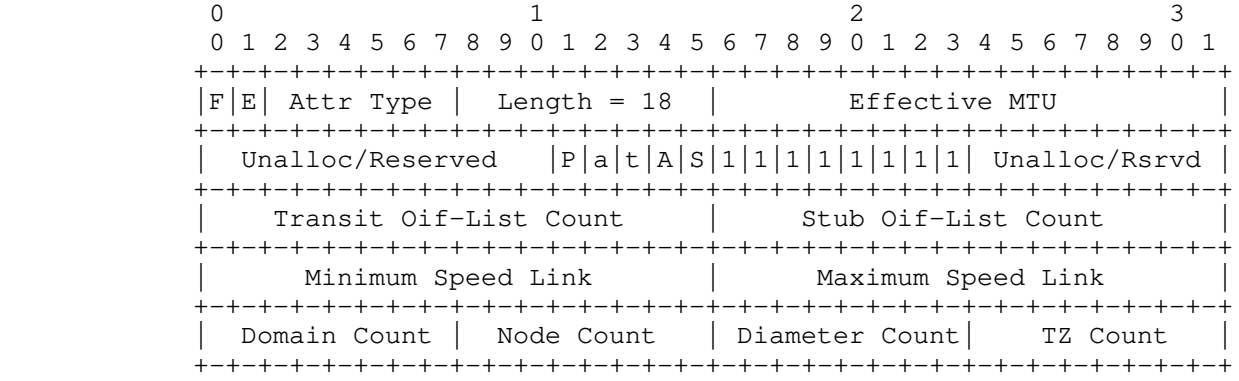
We will here give a few examples to illustrate the use of flags and options.

A minimum size message has no option flags set, and looks like this:

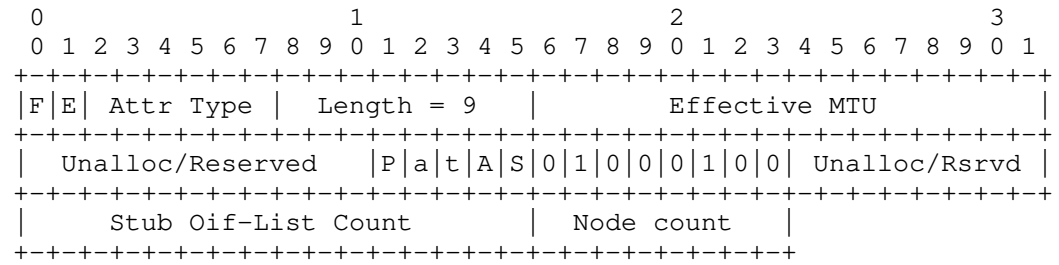


A message containing all the options defined in this document would look like this:

```
          <figure>
          <preamble></preamble>
          <artwork><![CDATA[
```



A message containing only Stub Oif-List Count and Node Count would look like this:


0 1 2 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
F E Attr Type
Length = 9
Effective MTU
Unalloc/Reserved
P a t A S 0 1 0 0 0 1 0 0
Unalloc/Rsrvd
Stub Oif-List Count
Node count

5. How to use Pop-Count Encoding

A router supporting this draft MUST include PIM Join Attribute TLV in its PIM Hellos. See [RFC5384] and [HELLO] for details.

It is very important to note that any changes to the values maintained in this draft MUST NOT trigger a new Join/Prune message. Due to the periodic nature of PIM, the values can be accurately obtained at 1 minute intervals (or whatever Join/Prune interval used).

When a router removes a link from an oif-list, it must be able to reevaluate the values that it will advertise upstream. This happens when an oif-list entry is timed out or a Prune is received.

It is recommended that the Join Attribute defined in this draft be used for entries in the join-list part of the Join/Prune message. If the new encoding is used in the prune-list or an Assert message, an implementation must ignore them but still process the Prune as if it was in the original encoding described in [RFC4601].

It is also recommended that join suppression be disabled on a LAN when Pop-Count is used.

6. Implementation Approaches

An implementation can decide how the accounting attributes are maintained. The values can be stored as part of the multicast route data structure by combining the local information it has with the joined information on a per oif basis. So when it is time to send a Join/Prune message, the values stored in the multicast route can be copied to the message.

Or, an implementation could store the accounting values per oif and when a Join/Prune message is sent, it can combine the oifs with its local information. Then the combined information can be copied to the message.

When a downstream joiner stops joining, accounting values cached must be evaluated. There are two approaches which can be taken. One is to keep values learned from each joiner so when the joiner goes away the count/max/min values are known and the combined value can be adjusted. The other approach is to set the value to 0 for the oif, and then start accumulating new values as subsequent Joins are received.

The same issue arises when an oif is removed from the oif-list. Keeping per-oif values allows you to adjust the per-route values when an oif goes away. Or, alternatively, a delay for reporting the new set a values from the route can occur while all oif values are zeroed (where accumulation of new values from subsequent Joins cause re-population of values and a new max/min/ count can be reevaluated for the route).

It is recommended that when triggered Join/Prune messages are sent by a downstream router, that the accounting information not be included in the message. This way when convergence is important, avoiding the processing time to build an accounting record in a downstream router and processing time to parse the message in the upstream router will help reduce convergence time. An upstream router should not interpret a Join/Prune message received with no accounting data to mean clearing or resetting what accounting data it has cached.

7. Caveats

This draft requires each router on a multicast distribution tree to support this draft or else the accounting attributes for the tree will not be known.

However, if there are a contiguous set of routers downstream in the distribution tree, they can maintain accounting information for the sub-tree.

If there are a set of contiguous routers supporting this draft upstream on the multicast distribution tree, accounting information will be available but it will not represent an accurate assessment of the entire tree. Also, it will not be clear for how much of the distribution tree the accounting information covers.

8. IANA Considerations

A new PIM Hello Option type, 29, has been assigned. See [HELLO] for details.

A new PIM Join Attribute type needs to be assigned. 2 is proposed in this draft.

9. Security Considerations

There are no security considerations for this design other than what is already in the main PIM specification [RFC4601].

10. Acknowledgments

The authors would like to thank John Zwiebel, Amit Jain, and Clayton Wagar for their review comments on the initial versions of this draft.

11. References

11.1. 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.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

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

11.2. Informative References

[AMT] Thaler, D., Talwar, M., Aggarwal, A., Vicisano, L., and T. Pusateri, "Automatic IP Multicast Without Explicit Tunnels (AMT)", draft-ietf-mboned-auto-multicast-10.txt (work in progress), March 2010.

[HELLO] IANA, "PIM Hello Options", PIM-HELLO-OPTIONS per RFC4601 http://www.iana.org/assignments/pim-hello-options, March 2007.

Authors' Addresses

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

Email: dino@cisco.com

Greg Shepherd
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: shep@cisco.com

Yiqun Cai
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: ycai@cisco.com

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

Email: stig@cisco.com

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: September 5, 2011

D. Farinacci
IJ. Wijnands
S. Venaas
cisco Systems
M. Napierala
AT&T Labs

March 4, 2011

A Reliable Transport Mechanism for PIM
draft-ietf-pim-port-06.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 September 5, 2011.

Copyright Notice

Table of Contents

1. Introduction

The goals of this specification are:

- 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].

- The reliable transport mechanism will be used for Join-Prune message transmission only.

- 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:

- Changes to the PIM message formats as defined in [RFC4601].

- Provide support for automatic switching between the reliable transport mechanism and the regular PIM mechanism defined in [RFC4601]. Two routers that are PIM neighbors on a link will always use the reliable transport mechanism if and only if both have enabled support for the reliable transport mechanism.

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 for 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 are still used for PIM neighbors that are not PORT enabled.

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. Also, any received native Join/Prune messages from that neighbor are discarded, even when the connection is down.

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. If there also are native joins/ prunes from non-PORT neighbor, then one can maintain timers on the

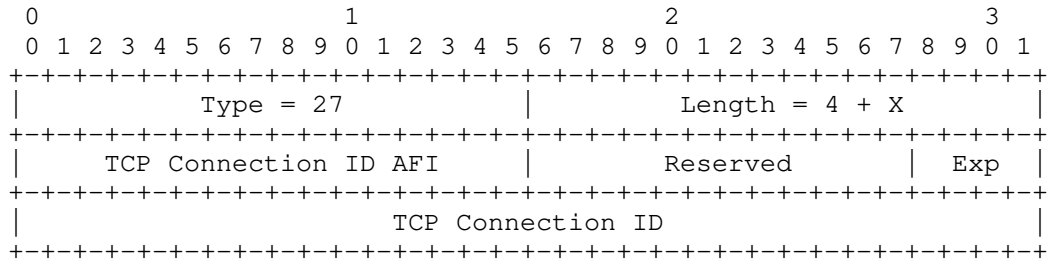
outgoing interface entry as usual, while at the same time keep track of each of the downstream PORT joins/prunes.

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. Each message is contained in a PORT message. See section Section 5 for details on the PORT message.

3. 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.

All Hello messages containing the PIM-over-TCP Capable hello option, MUST also contain the Interface ID hello option, see section Section 3.3.

Implementations MAY provide a configuration option to enable or disable PORT functionality. It is RECOMMENDED that this capability be disabled by default.

Length: Length in bytes for the value part of the Type/Length/Value encoding; where X is the number of bytes that make up the Connection ID field. X is 4 when AFI of value 1 (IPv4) is used, 16 when AFI of value 2 (IPv6) is used, and 0 if AFI of value 0 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 document is used to obtain the addresses used to establish the TCP connection.

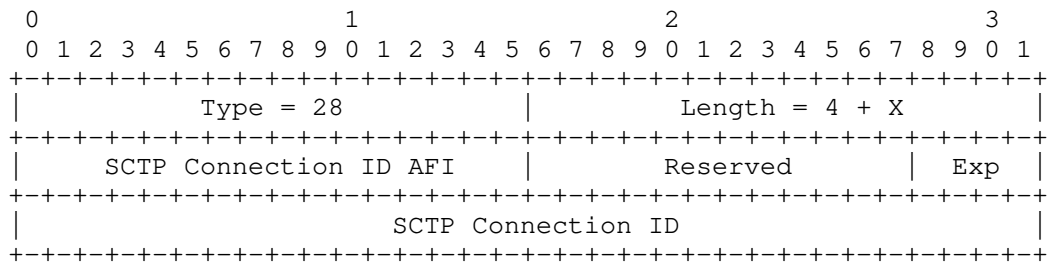
Reserved: Set to zero on transmission and ignored on receipt.

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.

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.

All Hello messages containing the PIM-over-SCTP Capable hello option, MUST also contain the Interface ID hello option, see section Section 3.3.

Implementations MAY provide a configuration option to enable or disable PORT functionality. It is RECOMMENDED that this capability be disabled by default.

Length: Length in bytes for the value part of the Type/Length/Value encoding; where X is the number of bytes that make up the Connection ID field. X is 4 when AFI of value 1 (IPv4) is used, 16 when AFI of value 2 (IPv6) is used, and 0 if AFI of value 0 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 document 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.

3.3. Interface ID

All Hello messages containing PIM-over-TCP Capable or PIM-over-SCTP Capable hello options, MUST also contain the Interface ID hello option [I-D.gulrajani-pim-hello-intid].

The 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 to convey the interface from Join/Prune message sender to Join/Prune message receiver. The value of the Interface ID hello option in Hellos sent on an interface, must be the same as the Interface ID value in all PORT Join/Prune messages sent to a PIM neighbor on that interface.

The Interface ID need only uniquely identify an interface of a router, it does not need to identify which router the interface belongs to. This means that the Router ID part of the Interface ID MAY be 0. For details on the Router ID and the value 0, see [I-D.gulrajani-pim-hello-intid].

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 4.5. 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 created before the connection was established (or reestablished) that is not refreshed, MUST be left to expire and be deleted. When the non-refreshed state has expired and been deleted, the two neighbors will be in sync.

It is possible that a router starts sending Hello messages with a new Connection ID, e.g. due to configuration changes. One MUST always use the last announced and last seen Connection IDs. A connection is identified by the local Connection ID (the one we are announcing on a particular interface), and the remote Connection ID (the one we are receiving from a neighbor on the same interface). When either the local or remote ID changes, the Connection ID pair we need a connection for changes. There may be an existing connection with the same pair, in which case we will share that connection. Or a new connection may need to be established. Note that for link-local addresses, the interface should be regarded as part of the ID, so that connection sharing is not attempted when the same link-local addresses are seen on different interfaces.

When a Connection ID changes, if the previously used connection is not needed (there are no other PIM neighborships using the same Connection ID pair), both peers MUST attempt a graceful shutdown of the connection. Next (even if the old connection is still needed), they MUST, unless a connection already exists with the new Connection ID pair, immediately or on-demand attempt to establish a new connection with the new Connection ID pair.

Normally the Interface ID would not change while a connection is up. However, if it does, it should not affect the connection. It just means that when subsequent PORT join/prune messages are received, they should be matched against the last seen Interface ID.

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 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. Connection Security

TCP/SCTP packets MUST be sent with a TTL/Hop Limit of 255 to facilitate enabling of the Generalized TTL Security Mechanism (GTSM) [RFC5082]. Implementations SHOULD provide a configuration option to enable the GTSM check. This means checking that inbound packets from directly connected neighbors have a TTL/Hop Limit of 255, but MAY also allow for a different TTL/Hop Limit threshold to check that the sender is within a certain number of router hops. The GTSM check SHOULD be disabled by default.

Implementations SHOULD support the TCP Authentication Option (TCP-AO) [RFC5925].

4.2. 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.

One can detect that a PORT connection is not working by regularly sending PORT messages. E.g., for TCP the connection will be reset if no TCP ACKs are received after a few retries. PORT in itself does not require any periodic signaling. PORT Join/Prune messages are only sent when there is a state change. If the state changes are not frequent enough, a PORT Keep-Alive message can be sent instead. E.g. if an implementation wants to send a PORT message, to check that the connection is working, at least every 60 seconds, then whenever there is 60 seconds since the previous message, a Keep-Alive message could be sent. If there were less than 60 seconds between each Join/Prune, no Keep-Alive messages would be needed. Implementations SHOULD support the use of PORT Keep-Alive messages. It is RECOMMENDED that a configuration option is available to network administrators to enable it when needed. Note that Keep-Alives can be used by a peer, independently of whether the other peer supports it.

The mechanism above relies on the connection eventually going down

when we don't get any ACKs for the data we send. A quicker and more reliable way of detecting that a connection is not working, is to send regular PORT messages, and have our peer take down the connection if it doesn't receive them. This can be done by sending Keep-alive messages with a non-zero holdtime value. If the last received Keep-alive message had a non-zero holdtime, one tears down the connection if the time measured in seconds since the last processed PORT message exceeds the specified holdtime.

Implementations SHOULD support Keep-Alive messages. An implementation that supports Keep-Alive messages acts as follows when processing a received PORT message. When processing a Keep-Alive message with a non-zero Holdtime value, it MUST set a timer to the value. We call this timer Connection Expiry Timer (CET). If the CET is already running, it MUST be reset to the new value. When processing a Keep-Alive message with a zero Holdtime value, the CET MUST be stopped if running. When processing a PORT message other than Keep-Alive, the CET MUST be reset to the last received Holdtime value if running. If the CET is not running, no action is taken. If the CET expires, the connection SHOULD be shut down.

It is possible that a router receives Join/Prune messages for an interface/link that is down. As long as the neighbor has not expired, it is RECOMMENDED processing those messages as usual. If they are ignored, then the router SHOULD ensure it gets a full update for that interface when it comes back up. This can be done by changing the GenID (Generation Identifier, see [RFC4601]), or by terminating and reestablishing the connection.

If a PORT neighbor changes its GenID and a connection is established or attempting to be established, the local side should generally tear down the connection and do as described in Section 4.3. However, if the connection is shared by multiple interfaces and the GenID changes only for one of them, the local side SHOULD simply send a full update, similar to other cases when a GenID changes for an upstream neighbor.

4.3. Actions When a Connection Goes Down

A connection may go down for a variety of reasons. It may be due to an error condition, or a configuration change. A connection SHOULD be shut down as soon as there are no more PIM neighbors 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. This may happen when a new connection ID is configured, PORT is disabled, or a PIM neighbor expires.

If a PIM neighbor expires, one should free connection state and downstream oif-list state for the neighbor. A downstream router, when an upstream neighboring router has expired, will simply update the RPF neighbor for the corresponding state to a new neighbor where it would trigger Join/Prune messages. This behavior is according to [RFC4601] where also the term RPF neighbor is defined. It is required of a PIM router to clear its neighbor table for a neighbor who has timed out due to neighbor holdtime expiration.

When a connection is no longer available between two PORT enabled PIM neighbors, they MUST immediately, or on-demand, try to reestablish the connection following the normal rules for connection establishment. The neighbors MUST also start expiry timers so that all oif-list state for the neighbor using the connection, gets expired after JP_HOLDTIME, unless it later gets refreshed by receiving new Join/Prunes.

The value of JP_HOLDTIME is 215 seconds. This value is based on section 4.11 of [RFC4601] which says that JP_HoldTime should be 3.5 * t_periodic where the default for t_periodic is 60 seconds.

4.4. 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, or a previously PORT enabled neighbor no longer announces any of the PORT Hello options, one follows the rules in Section 4.3 for taking down connections and starting timers. Next, one should trigger a full state update similar to what would be done if the GenID changed in Datagram Mode. This means sending joins for any state where we switched from PORT to Datagram Mode for the upstream neighbor.

4.5. 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 a full mesh of 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.

4.6. 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.

4.7. Avoiding a Pair of TCP Connections between Neighbors

To ensure that there is only one TCP connection between a pair of PIM neighbors, the following set of rules must be followed. Note that this section applies only to TCP, for SCTP this is not an issue. 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:

- 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.)

- 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.

- 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:

- 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.)

- 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. PORT Message Definition

It may be desirable for scaling purposes to allow Join/Prune messages from different PIM protocol families 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 PORT message format. This will provide both record boundary and demux points when sending over a stream protocol like TCP/SCTP.

A PORT message may contain PORT options, see Section 5.3. We will define two PORT options for carrying PIM Join/Prune messages. One for IPv4 and one for IPv6. For each PIM Join/Prune message to be sent over the Transport connection, we send a PORT Join/Prune message containing exactly one such option.

Each PORT message will have the Type/Length/Value format. 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.

PORT 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.

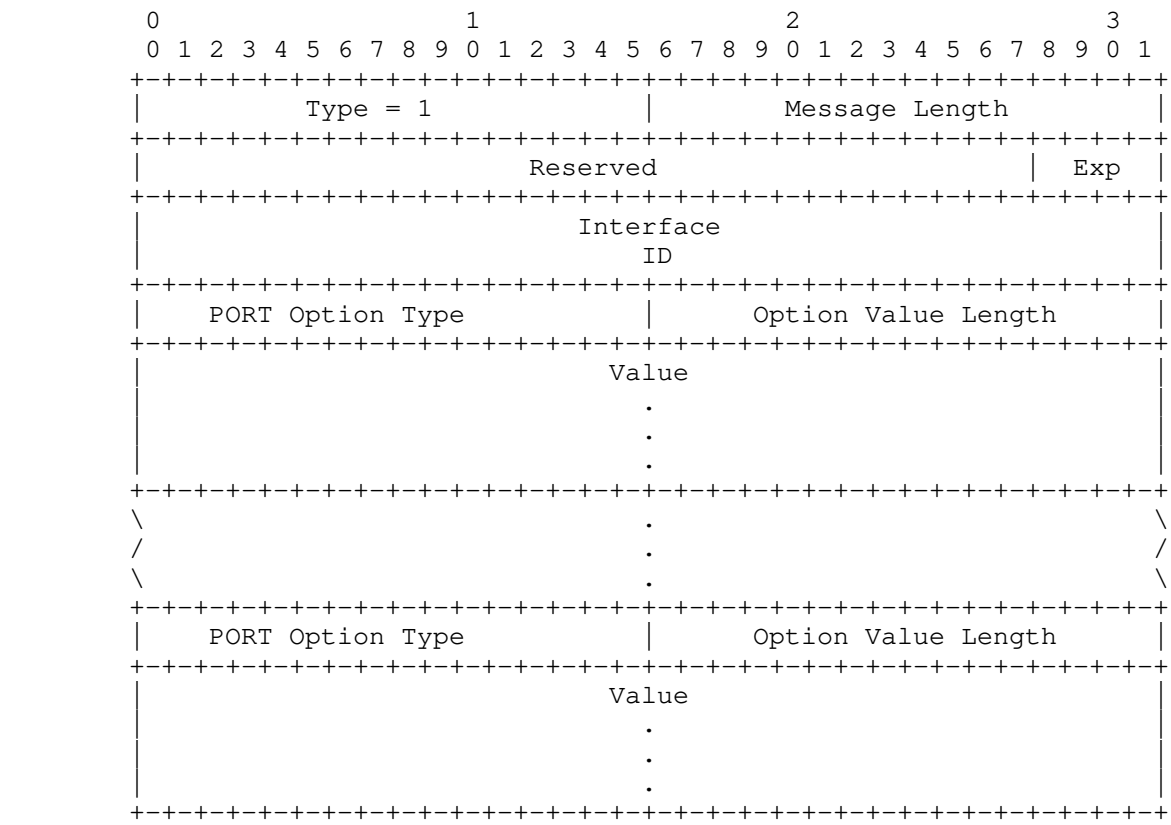
PORT messages are error checked. This includes a bad PIM checksum, illegal type fields, illegal addresses or a truncated message. If any parsing errors occur in a PORT message, it is skipped, and we proceed to any following PORT messages.

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

This document defines two message types.

5.1. PORT Join/Prune Message

PORT Join/Prune Message



The PORT Join/Prune Message is used for sending a PIM Join/Prune.

Message Length: Length in bytes for the value part of the Type/ Length/Value encoding. If no PORT Options were included, the length would be 12. If n PORT Options with Option Value lengths L1, L2, ..., Ln are included, the message length will be 12 + 4*n + L1 + L2 + ... + Ln.

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

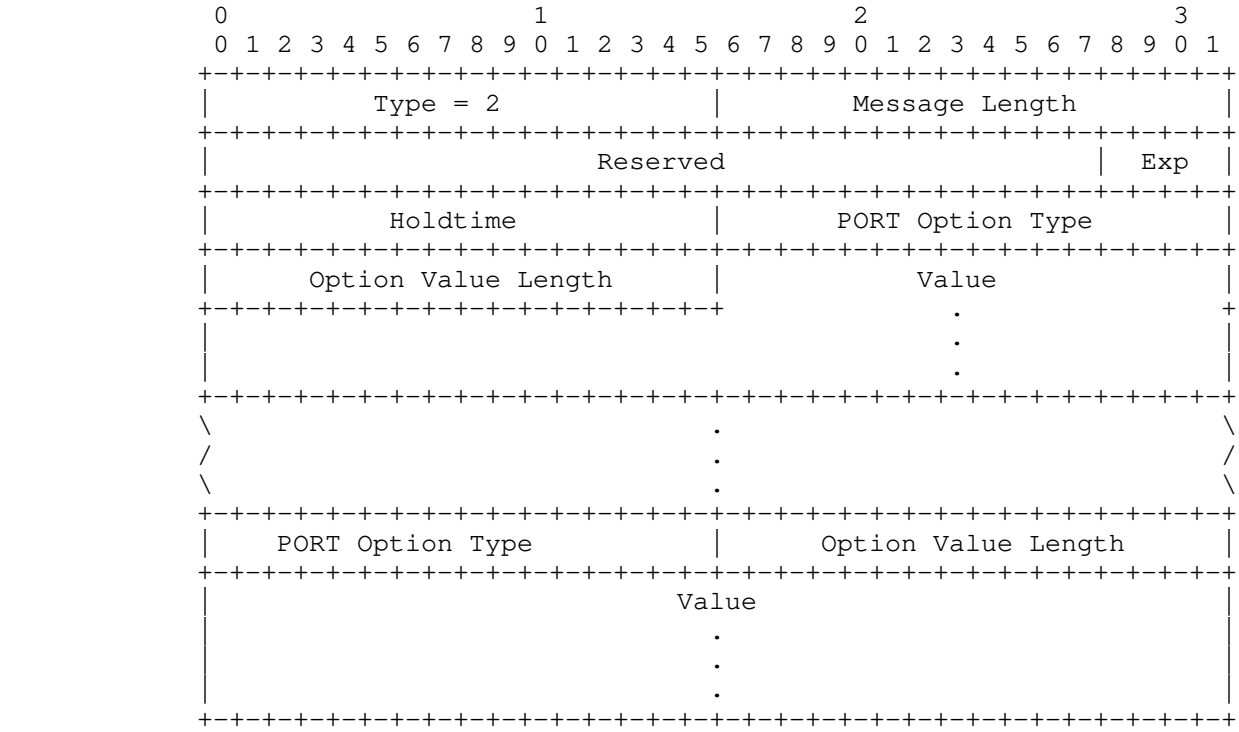
Exp: For experimental use [RFC3692].

Interface ID: This is the Interface ID of the Interface ID Hello option contained in the PIM Hello messages the PIM router is sending to the PIM neighbor. It indicates to the PIM neighbor what interface to associate the Join/Prune with. The Interface ID allows us to do connection sharing.

PORT Options: The message MUST contain exactly one PIM Join/Prune Port Option, either one PIM IPv4 Join/Prune or one PIM IPv6 Join/Prune. It MUST NOT contain both. It MAY contain additional options not defined in this document. A router receiving a PORT Join/Prune message containing unknown options MUST ignore the entire PORT message. See Section 5.3 for option definitions.

5.2. PORT Keep-alive Message

PORT Keep-alive Message



The PORT Keep-alive Message is used to regularly send PORT messages to verify that a connection is alive. They are used when other PORT messages are not sent at the desired frequency.

Message Length: Length in bytes for the value part of the Type/ Length/Value encoding. If no PORT Options were included, the length would be 6. If n PORT Options with Option Value lengths L1, L2, ..., Ln are included, the message length will be 6 + 4*n + L1 + L2 + ... + Ln.

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

Exp: For experimental use [RFC3692].

Holdtime: This specifies a holdtime in seconds for the connection. A non-zero value means that the connection SHOULD be gracefully shut down if no further PORT messages are received within the specified time. This is measured on the receiving side by measuring the time from one PORT message has been processed until the next has been processed. Note that this is done for any PORT message, not just keep-alive messages. A hold time of 0 disables the keep-alive mechanism.

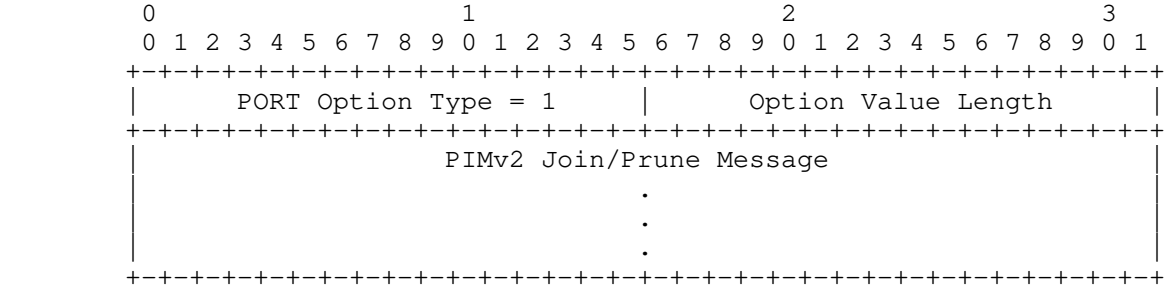
PORT Options: A keep-alive message MUST NOT contain any of the options defined in this document. It MAY contain other options not defined in this document. Unknown options MUST be ignored. See Section 5.3 for option definitions.

5.3. PORT Options

Each PORT Option is a TLV. The type is 16 bits. PORT Option types are assigned by IANA, except the range 61440 - 65535 which is for experimental use [RFC3692]. The length specifies the length of the value in bytes. Below are the two options defined in this document.

PIM IPv4 Join/Prune Option

PIM IPv4 Join/Prune Option Format



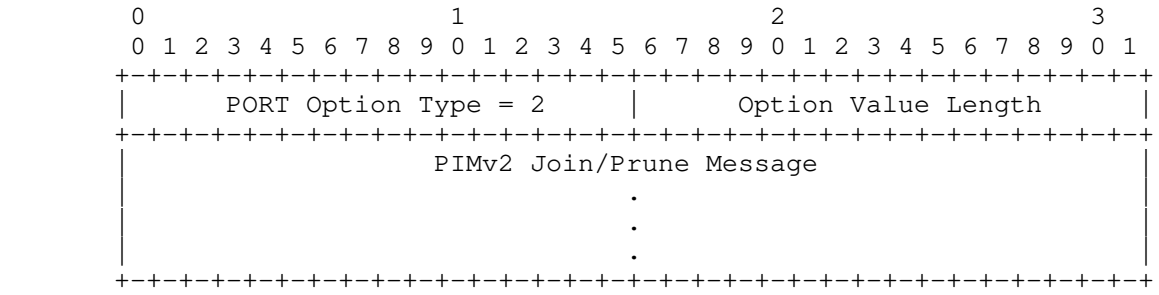
The IPv4 Join/Prune Option is used to carry a PIMv2 Join/Prune message that has all IPv4 encoded addresses in the PIM payload.

Option Value Length: The number of bytes that make up the PIMv2 Join/Prune message.

PIMv2 Join/Prune Message: PIMv2 Join/Prune message and payload with no IP header in front of it.

PIM IPv6 Join/Prune Option

PIM IPv6 Join/Prune Option Format



The IPv6 Join/Prune Option is used to carry a PIMv2 Join/Prune message that has all IPv6 encoded addresses in the PIM payload.

Option Value Length: The number of bytes that make up the PIMv2 Join/Prune message.

PIMv2 Join/Prune Message: PIMv2 Join/Prune message and payload with no IP header in front of it.

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, as specified [RFC4601], if no Join is sent to override this Prune before the expiration of the Override Timer, the upstream router concludes that no non-ET neighbor is interested. If no ET neighbors are interested, the interface can be removed from the oif-list. When an 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,rpt) 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,rpt)-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,rpt)-prunes, and if a non-ET neighbor has sent a (*,G)-join, whether there also is non-ET (S,G,rpt)-prune state.

7. Multiple Address-Family Support

To allow for efficient use of router resources, one can mux Join/ Prune messages of different address families on the same Transport connections. There are two ways this can be accomplished, one using a common message format over a TCP connection and the other using multiple streams over a single SCTP connection.

Using the common message format described previously in this specification, using different PORT options, both IPv4 and IPv6 based Join/Prune messages can be encoded within the same Transport connection.

When using SCTP multi-streaming, the common message format is still used to convey address family information but an SCTP association is used, on a per-family basis, to send data concurrently for multiple families. 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

TCP connections can be authenticated using TCP-AO [RFC5925]. When using SCTP, [RFC4895] can be used for authentication on a per SCTP association basis. Also GTSM [RFC5082] can be used to help prevent spoofing.

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.

10.1. PORT Hello Options

In the Protocol Independent Multicast (PIM) Hello Options registry, the following options are needed for PORT.

Value	Length	Name	Reference
27	Variable	PIM-over-TCP Capable	this document
28	Variable	PIM-over-SCTP Capable	this document

10.2. PORT Message Type Registry

A registry for PORT message types is requested. The message type is a 16-bit integer, with values from 0 to 65535. An RFC is required for assignments in the range 0 - 61439. This document defines one PORT message type. Type 1, PORT Join/Prune Message. The type range 61440 - 65535 is for experimental use [RFC3692].

The initial content of the registry should be as follows:

Type	Name	Reference
0	Reserved	this document
1	Join/Prune	this document
2	Keep-alive Message	this document
3-61439	Unassigned	
61440-65535	Experimental	this document

10.3. PORT Option Type Registry

A registry for PORT option types is requested. The option type is a 16-bit integer, with values from 0 to 65535. An RFC is required for assignments in the range 0 - 61439. This document defines two PORT option types. Type 1, PIM IPv4 Join/Prune Message; and Type 2, PIM IPv6 Join/Prune Message. The type range 61440 - 65535 is for experimental use [RFC3692].

The initial content of the registry should be as follows:

Type	Name	Reference
0	Reserved	this document
1	PIM IPv4 Join/Prune Message	this document
2	PIM IPv6 Join/Prune Message	this document
3-61439	Unassigned	
61440-65535	Experimental	this document

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, Gorry Fairhurst, Sameer Gulrajani, Thomas Morin, Dimitri Papadimitriou, Bharat Joshi, Rishabh Parekh, Manav Bhatia and Pekka Savola.

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

[I-D.gulrajani-pim-hello-intid]
Gulrajani, S. and S. Venaas, "An Interface ID Hello Option for PIM", draft-gulrajani-pim-hello-intid-00 (work in progress), February 2011.

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[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.

[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.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

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