

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 22, 2013

J. Arango  
S. Venaas  
I. Kouvelas  
Cisco Systems  
February 18, 2013

PIM Join Attributes for LISP Environments  
draft-arango-pim-join-attributes-for-lisp-00.txt

## Abstract

This document defines two PIM Join/Prune attributes that support the construction of multicast distribution trees where the root and receivers are located in different LISP sites. These attributes allow the receiver site to select between unicast and multicast transport and to convey the receiver RLOC address to the control plane of the root xTR.

## 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 22, 2013.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Requirements Notation . . . . .	4
3. PIM Join/Prune Attributes . . . . .	5
4. The Transport Attribute . . . . .	6
4.1. Transport Attribute Format . . . . .	6
4.2. Using the Transport Attribute . . . . .	6
5. Receiver RLOC Attribute . . . . .	8
5.1. Receiver RLOC Attribute Format . . . . .	8
5.2. Using the Receiver RLOC Attribute . . . . .	9
6. Security Considerations . . . . .	10
7. IANA Considerations . . . . .	11
8. Normative References . . . . .	12
Authors' Addresses . . . . .	13

## 1. Introduction

The construction of multicast distribution trees where the root and receivers are located in different LISP sites [RFC6830] is defined in [RFC6831]. Creation of (root-EID,G) state in the root site requires that unicast LISP-encapsulated Join/Prune messages be sent from an xTR on the receiver site to an xTR on the root site.

[RFC6831] specifies that (root-EID,G) data packets are to be LISP-encapsulated into (root-RLOC,G) multicast packets. However, a wide deployment of multicast connectivity between LISP sites is unlikely to happen any time soon. In fact, some implementations are initially focusing on unicast transport with head-end replication between root and receiver sites.

The unicast LISP-encapsulated Join/Prune message specifies the (root-EID,G) state that needs to be established in the root site, but conveys nothing about the receivers capability or desire to use multicast as the underlying transport. This document specifies a Join/Prune attribute that allows the receiver to select the desired transport.

Knowledge of the receiver RLOC is also essential to the control plane of the root xTR. It determines the downstream destination for unicast head-end replication and identifies the receiver xTR that needs to be notified should the root of the distribution tree move to another site.

The outer source address field of the encapsulated Join/Prune message contains an RLOC address of the receiver xTR. This source address is message to the root xTR RLOC destination. Due to policy and load balancing considerations, the selected source address may not be the RLOC on which the receiver site wishes to receive a particular flow. This document specifies a Join/Prune attribute that conveys the appropriate receiver RLOC address to the control plane of the root xTR.

## 2. 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].

### 3. PIM Join/Prune Attributes

PIM Join/Prune attributes are defined in [RFC5384] by introducing a new Encoded-Source type that, in addition to the Join/Prune source, can carry multiple type-length-value (TLV) attributes. These attributes apply to the individual Join/Prune sources on which they are stored.

The attributes defined in this document conform to the format of the encoding type defined in [RFC5384]. The attributes would typically be the same for all the sources in the Join/Prune message. Hence we RECOMMEND using the hierarchical Join/Prune attribute scheme defined in [I-D.venaas-pim-hierarchicaljoinattr]. This hierarchical system allows attributes to be conveyed on the Upstream Neighbor Address field, thus enabling the efficient application of a single attribute instance to all the sources in the Join/Prune message.

LISP xTRs do not exchange PIM Hello Messages and hence no Hello option is defined to negotiate support for these attributes. Systems that support unicast head-end replication are assumed to support these attributes.

#### 4. The Transport Attribute

It is essential that a mechanism be provided by which the desired transport can be conveyed by receiver sites. Root sites with multicast connectivity will want to leverage multicast replication. However, not all receiver sites can be expected to have multicast connectivity. It is thus desirable that root sites be prepared to support (root-EID,G) state with a mixture of multicast and unicast output state. This document specifies a Join/Prune attribute that allows the receiver to select the desired underlying transport.

##### 4.1. Transport Attribute Format

```

      0                               1                               2
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
      +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
      |F|E| Type = 5 | Length = 1 | Transport |
      +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

**F-bit:** The Transitive bit. Specifies whether the attribute is transitive or non-transitive. MUST be set to zero. This attribute is ALWAYS non-transitive.

**E-bit:** End-of-Attributes bit. Specifies whether this attribute is the last. Set to zero if there are more attributes. Set to 1 if this is the last attribute.

**Type:** The Transport Attribute type is 5.

**Length:** The length of the Transport Attribute value. MUST be set to 1.

**Transport:** The type of transport being requested. Set to 0 for multicast. Set to 1 for unicast.

##### 4.2. Using the Transport Attribute

Hierarchical Join/Prune attribute instances [I-D.venaas-pim-hierarchicaljoinattr] SHOULD be used when the same Transport Attribute is to be applied to all the sources within the Join/Prune message or all the sources within a group set. The root xTR MUST accept Transport Attributes in the Upstream Neighbor Encoded-Unicast address, Encoded-Group addresses, and Encoded-Source addresses.

There MUST NOT be more than one Transport Attribute within the same encoded address. If an encoded address has more than one instance of the attribute, the root xTR MUST discard all affected Join/Prune sources.

## 5. Receiver RLOC Attribute

The root xTR must know the receiver RLOC addresses of all receiver sites for a given (root-EID,G) so that it can perform unicast LISP-encapsulation of multicast data packets to each and every receiver site that has requested unicast head-end replication.

To support mobility of EIDs, the root xTR must keep track of ALL receiver RLOCs even when the corresponding downstream site has not requested unicast replication. The root xTR may detect that a local multicast source "root-EID" has moved to a remote LISP site. Under such circumstances LISP sends a SMR message to all receiver xTRs, prompting them to update their map cache. This is only possible if LISP can obtain from PIM the set of all receiver RLOCs that have active Join state for the root-EID.

The outer source address field of the encapsulated Join/Prune message contains an RLOC address of the receiver xTR. LISP xTRs, as edge devices, are commonly subject to URPF checks by the network providers on each core-facing interface. The source address for the encapsulation header must therefore be the RLOC of the core-facing interface used to physically transmit the encapsulated Join/Prune message. Due to policy and load balancing considerations, that may not be the RLOC on which the receiver site wishes to receive a particular flow. This document specifies a Join/Prune attribute that conveys the appropriate receiver RLOC address to the control plane of the root xTR.

### 5.1. Receiver RLOC Attribute Format

```

      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| Type = 6 | Length | Addr Family | Receiver RLOC
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+...
```

F-bit: The Transitive bit. Specifies whether this attribute is transitive or non-transitive. MUST be set to zero. This attribute is ALWAYS non-transitive.

E-bit: End-of-Attributes bit. Specifies whether this attribute is the last. Set to zero if there are more attributes. Set to 1 if this is the last attribute.



Type: The Receiver RLOC Attribute type is 6.

Length: The length in octets of the attribute value. MUST be set to the length in octets of the receiver RLOC address plus one octet to account for the Address Family field.

Addr Family: The PIM Address Family of the receiver RLOC as defined in [RFC4601].

Receiver RLOC: The RLOC address on which the receiver xTR wishes to receive the unicast-encapsulated flow.">

## 5.2. Using the Receiver RLOC Attribute

Hierarchical Join/Prune attribute instances [I-D.venaas-pim-hierarchicaljoinattr] SHOULD be used when the same Receiver RLOC attribute is to be applied to all the sources within the message or all the sources within a group set. The root xTR MUST accept Transport Attributes in the Upstream Neighbor Encoded-Unicast address, Encoded-Group addresses, and Encoded-Source addresses.

There MUST NOT be more than one Receiver RLOC Attribute within the same encoded address. If an encoded address has more than one instance of the attribute, the root xTR MUST discard all affected Join/Prune sources.

## 6. Security Considerations

Security of the Join Attribute is only guaranteed by the security of the PIM packet. The attributes specified herein do not enhance or diminish the privacy or authenticity of a Join/Prune message. A site that legitimately or maliciously sends and delivers a Join/Prune message to another site will equally be able to append these and any other attributes it wishes.

## 7. IANA Considerations

Two new PIM Join/Prune attribute types need to be assigned. Type 5 is being requested for the Transport Attribute. Type 6 is being requested for the Receiver RLOC Attribute.

## 8. Normative References

- [AFI] IANA, "Address Family Numbers",  
<http://www.iana.org/assignments/address-family-numbers>.
- [I-D.venaas-pim-hierarchicaljoinattr]  
Venaas, S., Kouvelas, I., and J. Arango, "Hierarchical Join/Prune Attributes",  
draft-venaas-pim-hierarchicaljoinattr-00 (work in progress), February 2013.
- [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.
- [RFC5384] Boers, A., Wijnands, I., and E. Rosen, "The Protocol Independent Multicast (PIM) Join Attribute Format", RFC 5384, November 2008.
- [RFC6830] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "The Locator/ID Separation Protocol (LISP)", RFC 6830, January 2013.
- [RFC6831] Farinacci, D., Meyer, D., Zwiebel, J., and S. Venaas, "The Locator/ID Separation Protocol (LISP) for Multicast Environments", RFC 6831, January 2013.

Authors' Addresses

Jesus Arango  
Cisco Systems  
170 Tasman Drive  
San Jose, CA 95134  
USA

Email: [jeearango@cisco.com](mailto:jeearango@cisco.com)

Stig Venaas  
Cisco Systems  
170 Tasman Drive  
San Jose, CA 95134  
USA

Email: [stig@cisco.com](mailto:stig@cisco.com)

Isidor Kouvelas  
Cisco Systems  
170 Tasman Drive  
San Jose, CA 95134  
USA

Email: [kouvelas@cisco.com](mailto:kouvelas@cisco.com)



PIM Working Group  
Internet-Draft  
Expires: August 29, 2013

H. Asaeda  
NICT  
S. Jeon  
Institute de Telecomunicacoes  
February 25, 2013

Multiple Upstream Interface Support for IGMP/MLD Proxy  
draft-asaeda-pim-mldproxy-multif-01

Abstract

This document describes the way of supporting multiple upstream interfaces for an IGMP/MLD proxy device. The proposed extension enables that an IGMP/MLD proxy device receives multicast packets through multiple upstream interfaces. The upstream interface is selected with manually configured supported address prefixes and interface priority value. A take-over operation switching from an inactive upstream interface to an active upstream interface is also considered.

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 29, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Terminology . . . . .	4
3. Per-Channel Load Balancing . . . . .	4
4. Candidate Upstream Interface Configuration . . . . .	5
4.1. Supported Address Prefix . . . . .	5
4.2. Interface Priority . . . . .	7
4.3. Default Interface . . . . .	7
5. IANA Considerations . . . . .	8
6. Security Considerations . . . . .	8
7. Normative References . . . . .	8
Authors' Addresses . . . . .	8



## 1. Introduction

The Internet Group Management Protocol (IGMP) [1][2] for IPv4 and the Multicast Listener Discovery Protocol (MLD) [3][2] for IPv6 are the standard protocols for hosts to initiate joining or leaving of multicast sessions. A proxy device performing IGMP/MLD-based forwarding (as known as IGMP/MLD proxy) [4] maintains multicast membership information by IGMP/MLD protocols on the downstream interfaces and sends IGMP/MLD membership report messages via the upstream interface to the upstream multicast routers when the membership information changes (e.g., by receiving solicited/unsolicited report messages). The proxy device forwards appropriate multicast packets received on its upstream interface to each downstream interface based on the downstream interface's subscriptions.

According to the specification of [4], an IGMP/MLD proxy has *\*a single\** upstream interface and one or more downstream interfaces. The multicast forwarding tree must be manually configured by designating upstream and downstream interfaces on an IGMP/MLD proxy device, and the root of the tree is expected to be connected to a wider multicast infrastructure. An IGMP/MLD proxy device hence performs the router portion of the IGMP or MLD protocol on its downstream interfaces, and the host portion of IGMP/MLD on its upstream interface. The proxy device must not perform the router portion of IGMP/MLD on its upstream interface.

On the other hand, there is a scenario in which an IGMP/MLD proxy device enables multiple upstream interfaces and receives multicast packets through these interfaces. For example, a proxy device having more than one interface may want to access to different networks, such as Internet and Intranet. Or, a proxy device having wired link (e.g., ethernet) and high-speed wireless link (e.g., WiMAX or LTE) may want to have the capability to connect to the Internet through both links. These proxy devices shall receive multicast packets from the different upstream interfaces and forward to the downstream interface(s).

This document adds the way to manually configure candidate upstream interfaces for an IGMP/MLD proxy device and select "one" single upstream interface from candidate upstream interfaces per session/channel. When the selected upstream interface is down or disabled, one of the other candidate upstream interfaces takes over the upstream interface (if configured). This enables "per-channel load balancing".

Note that this document only specifies the way to configure per-channel load balancing; it does not specify any intelligent

mechanism/algorithm (e.g., based on link or network condition/usage) or threshold value to select an upstream interface from candidate upstream interfaces to improve data reception quality. Also, an IGMP/MLD proxy device does not select multiple upstream interfaces for the same channels/sessions simultaneously; enabling redundant paths to receive duplicate packets via multiple upstream interfaces to improve data reception quality or robustness for a session/channel is out of scope of this document.

## 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 [5].

In addition, the following terms are used in this document.

Upstream interface (or selected upstream interface):

A proxy device's interface in the direction of the root of the multicast forwarding tree. An upstream interface is selected by either manual or automatic configuration.

Downstream interface:

Each of a proxy device's interfaces that is not in the direction of the root of the multicast forwarding tree.

Candidate upstream interface:

An interface that potentially becomes an upstream interface of the proxy device. Candidate upstream interfaces are manually set up on an IGMP/MLD proxy.

Supported address prefix:

The supported address prefix is the address prefix for which a candidate upstream interface supposes to be an upstream interface. The supported source address prefix and the supported multicast address prefix an IGMP/MLD proxy device can configure. The supported address prefix in this document means both source and multicast address prefixes, unless otherwise specified.

## 3. Per-Channel Load Balancing

An IGMP/MLD proxy device enables "per-channel load balancing" using multiple upstream interfaces to receive different multicast sessions/channel through the different upstream interfaces. Per-channel load balancing makes an IGMP/MLD proxy device select "one" single upstream interface from candidate upstream interfaces per session/channel,

based on the configurations, which will be described in Section 4.

If an IGMP proxy recognizes that an adjacent upstream router is not working, the selected upstream interface attached to that router can be taken over with the different candidate upstream interface. Or, if the selected upstream interface is going down, the proxy would switch from the inactive interface to the other active upstream interface. This "take-over operation" recursively examines the configurations of the candidate upstream interfaces (except the disabled interface) and decides a new upstream interface from them.

Whether the upstream router is active or not would be decided by checking a link condition or IGMP/MLD query message transmission. However, this document does not describe how an IGMP/MLD proxy can detect the upstream router's condition and when it takes that interface over the different candidate upstream interface.

The take-over operation is enabled by default. When it is disabled (by operation), even if no data comes from the selected upstream interface, the IGMP/MLD proxy device keeps using that interface as the upstream interface for the corresponding sessions/channels.

Per-channel load balancing does not implement duplicate packet reception from redundant paths using multiple upstream interfaces to improve data reception quality or robustness for a session/channel; therefore IGMP/MLD report messages containing the same IGMP/MLD records are not transmitted from different upstream interfaces simultaneously.

#### 4. Candidate Upstream Interface Configuration

Candidate upstream interfaces are the interfaces from which an IGMP/MLD proxy device selects as an upstream interface. They are manually enabled. The upstream interface selection is done based on "supported address prefix" and "interface priority" value.

##### 4.1. Supported Address Prefix

An IGMP/MLD proxy device MAY configure the "supported address prefix" for each candidate upstream interface. A proxy selects an upstream interface from its candidate upstream interfaces based on the configured supported address prefix. The supported address prefix is manually configured. The supported address prefix consists of the following information:

(source address prefix, multicast address prefix)

When the proxy device transmits an IGMP/MLD report message, it examines the source and multicast addresses in the IGMP/MLD records of the report message and transmits the appropriate IGMP/MLD report message(s) from the selected upstream interface(s) that are configured with the range of the supported source and multicast address prefixes.

The default values of both source and multicast address prefixes are a wildcard. If no address prefix value is configured on a candidate upstream interface, the default value is implicitly set up for the candidate upstream interface. The wildcard multicast address prefix is represented by the entire multicast address range (i.e., '224.0.0.0/4' for IPv4 or 'ff00::/8' for IPv6). The wildcard source address prefix is represented by any host. If the default value is set up on a candidate upstream interface, the decision whether the candidate upstream interface is selected as the upstream interface or not is made by the "interface priority" value described in Section 4.2.

The same address prefix may be configured on different candidate upstream interfaces. As well as the above-mentioned default configuration, when the same address prefix is configured on different candidate upstream interfaces, an upstream interface for that address prefix is selected based on each interface priority value described in Section 4.2.

For upstream interface selection, source address prefix takes priority over multicast address prefix. This avoids conflict of upstream interface selection. For example, consider the case that an IGMP/MLD proxy device has a configuration with source address prefix *S\_p* for the candidate upstream interface A and multicast address prefix *G\_p* for the candidate upstream interface B. When it deals with an IGMP/MLD record whose source address, let's say *S*, is in the range of *S\_p*, and whose multicast address, let's say *G*, is in the range of *G\_p*, the proxy device selects the candidate upstream interface A, which supports the source address prefix, as the upstream interface, and transmits the (*S*,*G*) record via the interface A.

Obviously, an IGMP/MLD proxy selects a candidate upstream interface having supported source and multicast address prefixes that include both source and multicast address, rather than the other one whose supported source and multicast address prefixes includes either source or multicast address.

#### 4.2. Interface Priority

An IGMP/MLD proxy device MAY configure the "interface priority" value for each candidate upstream interface. It is an integer value and manually configured. The default value of the interface priority is the lowest value.

The interface priority value effects only when the following conditions are satisfied.

- o None of the candidate upstream interfaces configure the supported address prefix.
- o Both source and multicast addresses are included in the supported address prefixes configured by more than one candidate upstream interface.
- o Neither source nor multicast address is included in the supported address prefixes configured by any of the candidate upstream interfaces.
- o The supported source address prefix is not configured or does not include the source address, but (on the other hand) the multicast address is included in the supported multicast address prefix configured by more than one candidate upstream interface.

In these conditions, the candidate upstream interface with the highest priority is chosen as the upstream interface.

#### 4.3. Default Interface

In the following conditions, the candidate upstream interface whose IPv4/v6 address is lowest is selected as the upstream interface for that session/channel.

- o None of the candidate upstream interfaces configure the supported address prefix and interface priority value.
- o Both source and multicast addresses are included in the supported address prefixes configured by more than one candidate upstream interfaces, and these candidate upstream interfaces' priorities are identical.
- o Neither source nor multicast address is included in the supported address prefixes configured by any of the candidate upstream interfaces, and all candidate upstream interfaces' priorities are identical.

- o The supported source address prefix is not configured or does not include the source address, and the multicast address is included in the supported multicast address prefix configured by more than one candidate upstream interface, yet these candidate upstream interfaces' priorities are identical.

## 5. IANA Considerations

This document has no actions for IANA.

## 6. Security Considerations

This document neither provides new functions nor modifies the standard functions defined in [1][3][2]. Therefore there is no additional security consideration provided for these protocols.

## 7. Normative References

- [1] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [2] 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.
- [3] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [4] 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.
- [5] Bradner, S., "Key words for use in RFCs to indicate requirement levels", RFC 2119, March 1997.

Authors' Addresses

Hitoshi Asaeda  
National Institute of Information and Communications Technology (NICT)  
Network Architecture Laboratory  
4-2-1 Nukui-Kitamachi  
Koganei, Tokyo 184-8795  
Japan

Email: asaeda@nict.go.jp

Seil Jeon  
Institute de Telecomunicacoes  
Campus Universitario de Santiago  
Aveiro 3810-193  
Portugal

Email: seiljeon@av.it.pt





Versions: 01

PIM WG  
Internet-Draft  
Intended status: Informational  
Expires: August 16, 2013

J. Asghar  
IJ. Wijnands  
S. Krishnaswamy  
A. Karan  
Cisco Systems  
V. Arya  
Directv, Inc.

February 19, 2013

Explicit RPF Vector  
draft-asghar-pim-explicit-rpf-vector-01

Abstract

This document describes a use of the Reverse Path Forwarding (RPF) Vector TLV as defined in [RPC 5496] to build multicast trees via an explicitly configured path sent in the PIM join.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire April 15, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this

document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Specification of Requirements . . . . .	3
3. Solution Requirements . . . . .	3
4. Use of the Explicit RPF Vector . . . . .	4
5. Explicit RPF Vector Attribute . . . . .	4
6. Conflicting RPF Vectors . . . . .	4
7. Explicit RPF Vector Attribute TLV Format . . . . .	4
8. IANA Considerations . . . . .	5
9. Security Considerations . . . . .	5
10. Acknowledgments . . . . .	6
11. Normative References . . . . .	6
Authors' Addresses . . . . .	7

## 1. Introduction

In some applications it might be useful to have a way to specify the explicit path along which the PIM join is propagated.

This document defines a new TLV in the PIM Join Attribute message [RFC5384] for specifying the explicit path.

The procedures in [RFC5496] define how a RPF vector can be used to influence the path selection in the absence of a route to the Source. However, the same procedures can be used to override a route to the Source when it exists. It is possible to include multiple RPF vectors in the stack where each router along the path will perform a unicast route lookup on the first vector in the attribute list. Once the router owning the address of the RPF vector is reached, following the procedures in [RFC5496], the RPF vector will be removed from the attribute list. This will result in a 'loosely' routed path based on the unicast reachability of the RPF vector(s). We call this loosely because we still depend on unicast routing reachability to the RPF Vector.

In some scenarios we don't want to rely on the unicast reachability to the RPF vector address and we want to build a path strictly based on the RPF vectors. In that case the RPF vector(s) represent a list of directly connected PIM neighbors along the path. For these vectors we MUST NOT do a unicast route lookup. We call these 'explicit' RPF vector addresses. If a router receiving an explicit RPF vector does not have a PIM neighbor matching the explicit RPF vector address it MUST NOT fall back to loosely routing the JOIN. Since the behavior of the explicit RPF vector differs from the loose RPF vector as defined [RFC5496], we're defining a new attribute called the explicit RPF Vector.

## 2. Specification of Requirements

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].

### 3. Solution Requirements

Some broadcast video transport networks use a multicast PIM live-live resiliency model for video delivery based on PIM SSM or PIM ASM. Live-Live implies using 2 active-active spatially diverse multicast trees to transport video flows from root to leaf multicast routers. The leaf multicast router receives 2 copies from the PIM multicast core and will replicate 1 copy towards the receivers [draft-mofrr-karan].

One of the main requirements of PIM live-live resiliency model is to ensure path-diversity of the active-active PIM trees in the core such that they do not intersect to avoid a single point of failure. IGP routed RPF paths of active-active PIM trees could be routed over the same transit router and create a single point of failure. It might be useful to have a way to specify the explicit path along which the PIM join is propagated.

How the explicit RPF vector stack is determined is outside the scope of this document. It may either be manually configured by the network operator or procedures may be implemented on the egress router to dynamically calculate the vector stack based on a link state database protocol, like OSPF or ISIS.

Due to the fact that the leaf router receives two copies of the multicast stream via two diverse paths, there is no need for PIM to repair the broken path immediately. It is up to the egress router to either wait for the broken path to be repaired or build a new explicit path using a new RPF vector stack. Which method is applied depends very much on how the vector stack was determined initially. Double failures are not considered and outside the scope of this document

#### 4. Use of the PIM Explicit RPF Vector

Figure 1 provides an example multicast join path R4->R3->R6->R5->R2->R1, where the multicast JOIN is explicitly routed to the source hop-by-hop using the explicit RPF vector list.

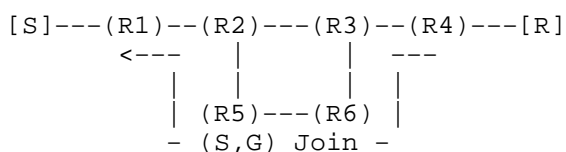


Figure 1

## 5. Explicit RPF Vector Attribute

This draft uses vector attribute 4 for specifying an explicit RPF vector.

## 6. Conflicting RPF Vectors

It is possible that a PIM router has multiple downstream neighbors. If for the same multicast route there is inconsistency between the Explicit RPF Vector stacks provided by the downstream PIM neighbor, the procedures as documented in RFC5384 section 3.3.3 apply.

## 7. Explicit RPF Vector Attribute TLV Format



#### Authors' Addresses

Javed Asghar  
Cisco Systems, Inc.  
725, Alder Drive  
Milpitas, CA 95035

Email: [jasghar@cisco.com](mailto:jasghar@cisco.com)

IJsbrand Wijnands  
Cisco Systems, Inc.  
De kleetlaan 6a  
Diegem 1831  
Belgium

EMail: [ice@cisco.com](mailto:ice@cisco.com)

Sowmya Krishnaswamy  
Cisco Systems, Inc.  
3750 Cisco Way  
San Jose, CA 95134

EMail: [sowkrish@cisco.com](mailto:sowkrish@cisco.com)

Apoorva Karan  
Cisco Systems, Inc.  
3750 Cisco Way  
San Jose, CA 95134

EMail: [apoorva@cisco.com](mailto:apoorva@cisco.com)

Vishal Arya  
DIRECTV Inc.  
2230 E Imperial Hwy  
El Segundo, CA 90245

Email: [varya@directv.com](mailto:varya@directv.com)

MULTIMOB Working Group  
INTERNET-DRAFT  
Intended Status: Proposed Standard  
Expires: April 18, 2013

Luis M. Contreras  
Telefonica I+D  
Carlos J. Bernardos  
Universidad Carlos III de Madrid  
Juan Carlos Zuniga  
InterDigital  
February 25, 2013

Extension of the MLD proxy functionality to support multiple  
upstream interfaces  
draft-contreras-multimob-multiple-upstreams-01

## Abstract

This document presents different scenarios of applicability for an MLD proxy running more than one upstream interface. Since those scenarios impose different requirements on the MLD proxy with multiple upstream interfaces, it is important to ensure that the proxy functionality addresses all of them for compatibility.

The purpose of this document is to define the requirements in an MLD proxy with multiple interfaces covering a variety of applicability scenarios, and to specify the proxy functionality to satisfy all of them.

## Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

## Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1	Introduction . . . . .	4
2.	Terminology . . . . .	4
3.	Problem statement . . . . .	4
4.	Scenarios of applicability . . . . .	7
4.1	Fixed network scenarios . . . . .	7
4.1.1	Multicast wholesale offer for residential services . . . . .	7
4.1.1.1	Requirements . . . . .	7
4.1.2	Multicast resiliency . . . . .	8
4.1.2.1	Requirements . . . . .	8
4.1.3	Load balancing for multicast traffic in the metro segment . . . . .	8
4.1.3.1	Requirements . . . . .	8
4.1.4	Summary of the requirements needed for mobile network scenarios . . . . .	9
4.2	Mobile network scenarios . . . . .	9
4.2.1	Applicability to multicast listener mobility . . . . .	10
4.2.1.1	Single MLD proxy instance on MAG . . . . .	10
4.2.1.1.1	Requirements . . . . .	10
4.2.1.2	Remote and local multicast subscription . . . . .	10
4.2.1.2.1	Requirements . . . . .	11
4.2.1.3	Dual subscription to multicast groups during handover . . . . .	11
4.2.1.3.1	Requirements . . . . .	12
4.2.2	Applicability to multicast source mobility . . . . .	12
4.2.2.1	Support of remote and direct subscription in basic source mobility . . . . .	12
4.2.2.1.1	Requirements . . . . .	13
4.2.2.2	Direct communication between source and listener associated with distinct LMAs but on the same MAG . . . . .	13

4.2.2.3.1	Requirements . . . . .	14
4.2.2.3	Route optimization support in source mobility for remote subscribers . . . . .	14
4.2.2.3.1	Requirements . . . . .	14
4.2.3	Summary of the requirements needed for mobile network scenarios . . . . .	15
5	Functional specification of an MLD proxy with multiple interfaces . . . . .	17
6	Security Considerations . . . . .	17
7	IANA Considerations . . . . .	17
8	Conclusions . . . . .	17
9	Acknowledgements . . . . .	17
10	References . . . . .	17
10.1	Normative References . . . . .	17
10.2	Informative References . . . . .	17
	Appendix A. Basic support for multicast listener with PMIPv6 . .	18
	Authors' Addresses . . . . .	20



## 1 Introduction

The aim of this document is to define the functionality that an MLD proxy with multiple upstream interfaces should have in order to support different scenarios of applicability in both fixed and mobile networks. This compatibility is needed in order to simplify node functionality and to ensure an easier deployment of multicast capabilities in all the use cases described in this document.

## 2. Terminology

This document uses the terminology defined in [3]. Specifically, the definition of Upstream and Downstream interfaces, which are reproduced here for completeness.

Upstream interface:

A proxy device's interface in the direction of the root of the tree. Also called the "Host interface".

Downstream interface:

Each of a proxy device's interfaces that is not in the direction of the root of the tree. Also called the "Router interfaces".

## 3. Problem statement

The concept of MLD proxy with several upstream interfaces has emerged as a way of optimizing (and in some cases enabling) service delivery scenarios where separate multicast service providers are reachable through the same access network infrastructure. Figure 1 presents the conceptual model under consideration.

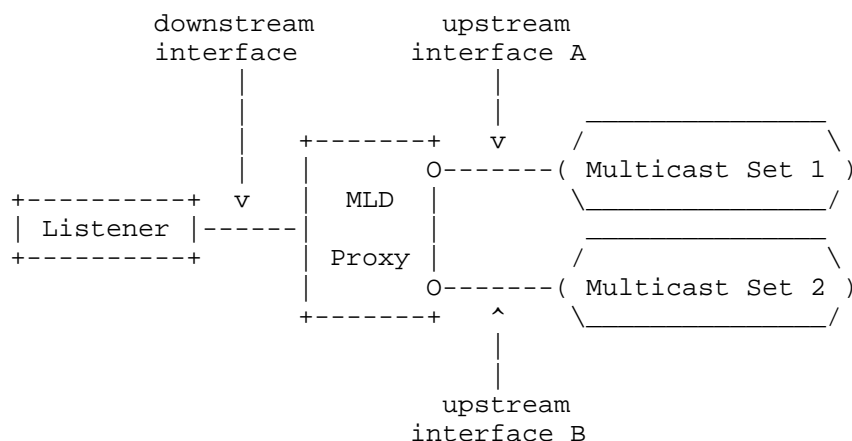


Figure 1. Concept of MLD proxy with multiple upstream interfaces

For illustrative purposes, two applications for fixed and mobile networks are here introduced. They will be elaborated later on the document.

In the case of fixed networks, multicast wholesale services in a competitive residential market require an efficient distribution of multicast traffic from different operators, i.e. the incumbent operator and a number of alternative ones, on the network infrastructure of the former. Existing proposals are based on the use of PIM routing from the metro network, and multicast traffic aggregation on the same tree. A different approach could be achieved with the use of an MLD proxy with multiple upstream interfaces, each of them pointing to a distinct multicast router in the metro border which is part of separated multicast trees deep in the network. Figure 2 graphically describes this scenario.

In the case of mobile networks, IP mobility services guarantee the continuity of the IP session while a Mobile Node (MN) changes its point of attachment. Proxy Mobile IPv6 (PMIPv6) [1] standardized a protocol that allows the network to manage the MN mobility without requiring specific support from the mobile terminal. The traffic to the MN is tunneled from the Home Network making use of two entities, one acting as mobility anchor, and the other as Mobility Access Gateway (MAG). Multicast support in PMIPv6 [2] implies the delivery of all the multicast traffic from the Home Network, via the mobility anchor. However, multicast routing optimization [4] could take advantage of an MLD proxy with multiple upstream interfaces by supporting the decision of subscribing a multicast content from the Home Network or from the local PMIPv6 domain if it is locally available. Figure 3 presents this scenario.

Informational text is provided in Appendix A summarizing how the basic solution for deploying multicast listener mobility with Proxy Mobile IPv6 works.

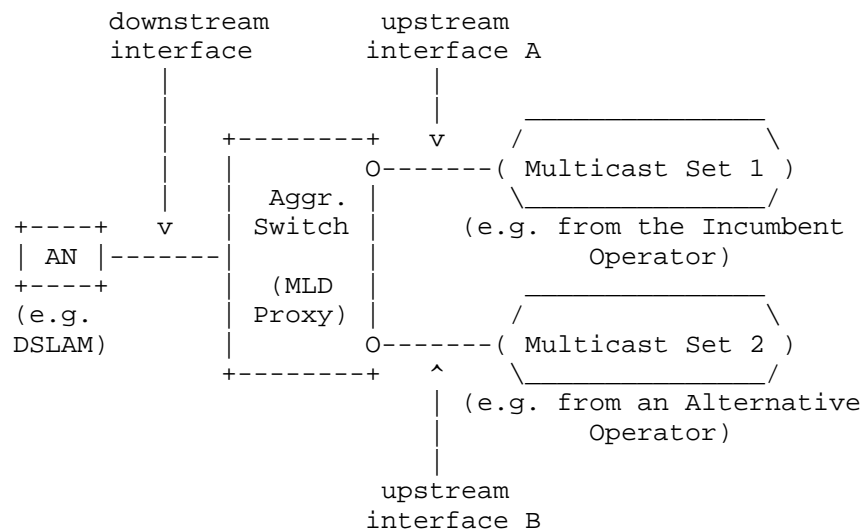


Figure 2. Example of usage of an MLD proxy with multiple upstream interfaces in a fixed network scenario

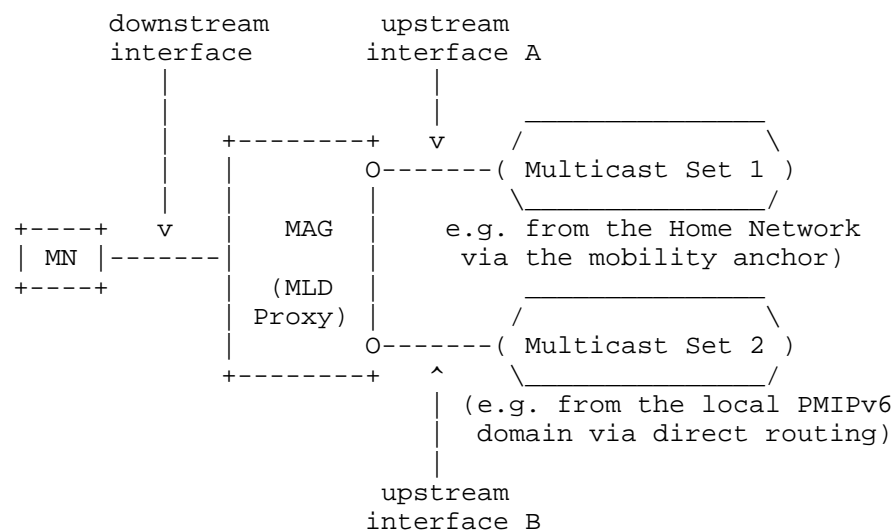


Figure 3. Example of usage of an MLD proxy with multiple upstream interfaces in a mobile network scenario

Since those scenarios can motivate distinct needs in terms of MLD proxy functionality, it is necessary to consider a comprehensive approach, looking at the possible scenarios, and establishing a minimum set of requirements which can allow the operation of a versatile MLD proxy with multiple upstream interfaces as a common entity to all of them (i.e., no different kinds of proxies depending on the scenario, but a common proxy applicable to all the potential scenarios).

#### 4. Scenarios of applicability

This section describes in detail a number of scenarios of applicability of an MLD proxy with multiple upstream interfaces in place. A number of requirements for the MLD proxy functionality are identified from those scenarios.

##### 4.1 Fixed network scenarios

Residential broadband users get access to multiple IP services through fixed network infrastructures. End user's equipment is connected to an access node, and the traffic of a number of access nodes is collected in aggregation switches.

For the multicast service, the use of an MLD proxy with multiple upstream interfaces in those switches can provide service flexibility in a lightweight and simpler manner if compared with PIM-routing based alternatives.

##### 4.1.1 Multicast wholesale offer for residential services

This scenario has been already introduced in the previous section, and can be seen in Figure 2. There are two different operators, the one operating the fixed network where the end user is connected (e.g., typically an incumbent operator), and the one providing the Internet service to the end user (e.g., an alternative Internet service provider). Both can offer multicast streams that can be subscribed by the end user, independently of which provider contributes with the content.

Note that it is assumed that both providers offer distinct multicast groups. However, more than one subscription to multicast channels of different providers could take place simultaneously.

##### 4.1.1.1 Requirements

- The MLD proxy should be able to deliver multicast control messages sent by the end user to the corresponding provider's multicast router.

- The MLD proxy should be able to deliver multicast control messages sent by each of the providers to the corresponding end user.

#### 4.1.2 Multicast resiliency

In current PIM-based solutions, the resiliency of the multicast distribution relays on the routing capabilities provided by protocols like PIM and VRRP. A simpler scheme could be achieved by implementing different upstream interfaces on MLD proxies, providing path diversity through the connection to distinct leaves of a given multicast tree.

It is assumed that only one of the upstream interfaces is active in receiving the multicast content, while the other is up and in standby for fast switching.

##### 4.1.2.1 Requirements

- The MLD proxy should be able to deliver multicast control messages sent by the end user to the corresponding active upstream interface.
- The MLD proxy should be able to deliver multicast control messages received in the active upstream to the end users, while ignoring the control messages of the standby upstream interface.
- The MLD proxy should be able of rapidly switching from the active to the standby upstream interface in case of network failure, transparently to the end user.

#### 4.1.3 Load balancing for multicast traffic in the metro segment

A single upstream interface in existing MLD proxy functionality typically forces the distribution of all the channels on the same path in the last segment of the network. Multiple upstream interfaces could naturally split the demand, alleviating the bandwidth requirements in the metro segment.

##### 4.1.3.1 Requirements

- The MLD proxy should be able to deliver multicast control messages sent by the end user to the corresponding multicast router which provides the channel of interest.
- The MLD proxy should be able to deliver multicast control messages sent by each of the multicast routers to the corresponding end user.
- The MLD proxy should be able to decide which upstream interface is selected for any new channel request according to defined criteria

(e.g., load balancing).

#### 4.1.4 Summary of the requirements needed for mobile network scenarios

Following the analysis above, a number of different requirements can be identified by the MLD proxy to support multiple upstream interfaces in fixed network scenarios. The following table summarizes these requirements.

	Fixed Network Scenarios		
Functionality	Multicast Wholesale	Multicast Resiliency	Load Balancing
Upstream Control Delivery	X	X	X
Downstr. Control Delivery	X	X	X
Active / Standby Upstream		X	
Upstr i/f selection per group			X
Upstr i/f selection all group		X	

Table I. Functionality needed on MLD proxy with multiple upstream interfaces per application scenario in fixed networks

#### 4.2 Mobile network scenarios

The mobile networks considered in this document are supposed to run PMIPv6 protocol for IP mobility management. A brief description of multicast provision in PMIPv6-based networks can be found in Appendix A.

The use of an MLD proxy supporting multiple upstream interfaces can improve the performance and the scalability of multicast-capable PMIPv6 domains.

#### 4.2.1 Applicability to multicast listener mobility

Three sub-cases can be identified for the multicast listener mobility.

##### 4.2.1.1 Single MLD proxy instance on MAG

The base solution for multicast service in PMIPv6 [2] assumes that any MN subscribed to multicast services receive the multicast traffic through the associated LMA, as in the unicast case. As standard MLD proxy functionality only supports one upstream interface, the MAG should implement several separated MLD proxy instances, one per LMA, in order to serve the multicast traffic to the MNs, according to any particular LMA-MN association.

A way of avoiding the multiplicity of MLD proxy instance in a MAG is to deploy a unique MLD proxy instance with multiple upstream interfaces, one per LMA, without any change in the multicast traffic distribution.

##### 4.2.1.1.1 Requirements

- The MLD proxy should be able of delivering the multicast control messages sent by the MNs to the associated LMA.
- The MLD proxy should be able of delivering the multicast control messages sent by each of the connected LMAs to the corresponding MN.
- The MLD proxy should be able of routing the multicast data coming from different LMAs to the corresponding MNs according to the MN to LMA association.
- The MLD proxy should be able of maintaining a 1:1 association between an MN and LMA (or downstream to upstream).

##### 4.2.1.2 Remote and local multicast subscription

This scenario has been already introduced in the previous section, and can be seen in Figure 3. Standard MLD proxy definition, with a unique upstream interface per proxy, does not allow the reception of multicast traffic from distinct upstream multicast routers. In other words, all the multicast traffic being sent to the MLD proxy in

downstream traverses a concrete, unique router before reaching the MAG. There are, however, situations where different multicast content could reach the MLD proxy through distinct next-hop routers.

For instance, the solution adopted to avoid the tunnel convergence problem in basic multicast PMIPv6 deployments [4] considers the possibility of subscription to a multicast source local to the PMIPv6 domain. In that situation, some multicast content will be accessed remotely, through the home network via the multicast tree mobility anchor, while some other multicast content will reach the proxy directly, via a local router in the domain.

#### 4.2.1.2.1 Requirements

- The MLD proxy should be able of delivering the multicast control messages sent by the MNs to the associated upstream interface based on the location of the source, remote or local, for a certain multicast group.
- The MLD proxy should be able of delivering the multicast control messages sent either local or remotely to the corresponding MNs.
- The MLD proxy should be able of routing the multicast data coming from different upstream interfaces to a certain MN according to the MN subscription, either local or remote. Note that it is assumed that a multicast group can be subscribed either locally or remotely, but not simultaneously. However more than one subscription could happen, being local or remote independently.
- The MLD proxy should be able of maintaining a 1:N association between an MN and the remote and local multicast router (or downstream to upstream).
- The MLD proxy should be able of switching between local or remote subscription for per multicast group according to specific configuration parameters (out of the scope of this document).

#### 4.2.1.3 Dual subscription to multicast groups during handover

In the event of an MN handover, once an MN moves from a previous MAG (pMAG) to a new MAG (nMAG), the nMAG needs to set up the multicast status for the incoming MN, and subscribe the multicast channels it was receiving before the handover event. The MN will then experience a certain delay until it receives again the subscribed content.

A generic solution is being defined in [5] to speed up the knowledge of the ongoing subscription by the nMAG. However, for the particular case that the underlying radio access technology supports layer-2



triggers (thus requiring extra capabilities on the mobile node), there could be inter-MAG cooperation for handover support if pMAG and nMAG are known in advance.

This could be the case, for instance for those contents not already arriving to the nMAG, where the nMAG temporally subscribes the multicast groups of the ongoing MN's subscription via the pMAG, while the multicast delivery tree among the nMAG and the mobility anchor is being established.

A similar approach is followed in [6] despite the solution proposed there differs from this approach (i.e., there is no consideration of an MLD proxy with multiple interfaces).

#### 4.2.1.3.1 Requirements

- The MLD proxy should be able of delivering the multicast control messages sent by the MNs to the associated upstream interface based on the handover specific moment, for a certain multicast group.
- The MLD proxy should be able of delivering the multicast control messages sent either from pMAG or the multicast anchor to the corresponding MNs, based on the handover specific moment.
- The MLD proxy should be able of handle the incoming packet flows from the two simultaneous upstream interfaces, in order to not duplicate traffic delivered on the point-to-point link to the MN.
- The MLD proxy should be able of maintaining a 1:N association between an MN and both the remote multicast router and the pMAG (or downstream to upstream).
- The MLD proxy should be able of switching between local or remote subscription for all the multicast groups (from pMAG to multicast anchor) according to specific configuration parameters (out of the scope of this document).

#### 4.2.2 Applicability to multicast source mobility

A couple of sub-cases can be identified for the multicast source mobility.

##### 4.2.2.1 Support of remote and direct subscription in basic source mobility

In the basic case of source mobility, the multicast source is connected to one of the downstream interfaces of an MLD proxy. According to the standard specification [3] every packet sent by the

multicast source will be forwarded towards the root of the multicast tree.

However, linked to the mobility listener problem, there could be the case of simultaneous remote subscribers, subscribing to the multicast content through the home network, and local subscribers, requesting the contents directly via a multicast router residing on the same PMIPv6 domain where the source is attached to.

Then, in order to provide the co-existence of both types of subscribers, an MLD proxy with two upstream interfaces could simultaneously serve all kind of multicast subscribers.

Basic source mobility is being defined in [7] but the solution proposed there does not allow simultaneous co-existence of remote and local subscribers (i.e., the content sent by the source is either distributed locally to a multicast router in the PMIPv6 domain, or remotely by using the bi-directional tunnel towards the mobility anchor, but not both simultaneously).

#### 4.2.2.1.1 Requirements

- The MLD proxy should be able of forwarding (replicating) the multicast content to both upstream interfaces, in case of simultaneous remote and local distribution.
- The MLD proxy should be able of handling control information incoming through any of the two upstream interfaces, providing the expected behavior for each of the multicast trees.
- The MLD proxy should be able of routing the multicast data towards different upstream interfaces for both remote and local subscriptions that could happen simultaneously.
- The MLD proxy should be able of maintaining a 1:N association between an MN and both the remote and local multicast router (or downstream to upstream).

#### 4.2.2.2 Direct communication between source and listener associated with distinct LMAs but on the same MAG

In a certain PMIPv6 domain can be MNs associated to distinct LMAs using the same MAG to get access to their corresponding home networks. For multicast communication, according to the base solution [2], each MN <-> LMA association implies a distinct MLD proxy instance to be invoked in the MAG.

In these conditions, when a mobile source is serving multicast content to a mobile listener, both attached to the same MAG but each of them associated to different LMAs, the multicast flow must traverse the PMIPv6 domain from the MAG to the LMA where the source maintains an association, then from that LMA to the LMA where the listener is associated to, and finally come back to the same MAG from where the flow departed. This routing is extremely inefficient.

An MLD proxy with multiple upstream interfaces avoids this behavior since it allows to invoke a unique MLD proxy instance in the MAG. In this case, the multicast source can directly communicate with the multicast listener, without need for delivering the multicast traffic to the LMAs.

#### 4.2.2.3.1 Requirements

- The MLD proxy should be able of forwarding (replicating) the multicast content to different upstream or downstream interfaces where subscribers are present.
- The MLD proxy should be able of handling control information incoming through any of the upstream or downstream interfaces requesting a multicast flow being injected in another downstream interface.
- The MLD proxy should be able of maintaining a 1:N association between an MN and any of the upstream or downstream interfaces demanding the multicast content.

#### 4.2.2.3 Route optimization support in source mobility for remote subscribers

Even in a scenario of remote subscription, there could be the case where both the source and the listener are attached to the same PMIPv6-Domain (for instance, no possibility of direct routing within the PMIPv6, or source and listener pertaining to distinct home networks). In this situation there is a possibility of route optimization if inter-MAG communication is enabled, in such a way that the listeners in the PMIPv6 domain are served through the tunnels between MAGs, while the rest of remote listeners are served through the mobility anchor.

A multi-upstream MLD proxy would allow the simultaneous delivery of traffic to such kind of remote listeners.

A similar route optimization approach is proposed in [8].

#### 4.2.2.3.1 Requirements

- The MLD proxy should be able of forwarding (replicating) the multicast content to both kinds of upstream interfaces, inter-MAG tunnel interfaces and MAG to mobility anchor tunnel interface.
- The MLD proxy should be able of handling control information incoming through any of the two types of upstream interfaces, providing the expected behavior for each of the multicast trees (e.g., no forwarding traffic on one inter-MAG link once there are not more listeners requesting the content).
- The MLD proxy should be able of routing the multicast data towards different upstream interfaces for both remote and route optimized subscriptions that could happen simultaneously.
- The MLD proxy should be able of maintaining a 1:N association between an MN and both the remote and local MAGs (or downstream to upstream).

#### 4.2.3 Summary of the requirements needed for mobile network scenarios

After the previous analysis, a number of different requirements can be identified by the MLD proxy to support multiple upstream interfaces in mobile network scenarios. The following table summarizes these requirements.

Functionality	Mobile Network Scenarios					
	Multicast Listener			Multicast Source		
	Single MLD Proxy	Remote & local subscr.	Dual subscr. in HO	Direct & remote subscr.	Listener & source on MAG	Route optimi.
Upstream Control Delivery	X	X	X	X	X	X
Downstr. Control Delivery	X	X	X		X	
Upstream Data Delivery				X		X
Downstr. Data Delivery	X	X	X		X	
1:1 MN to upstream assoc.	X					
1:N MN to upstream assoc.		X	X	X	X	X
Upstr i/f selection per group		X				
Upstr i/f selection all group			X			
Upstream traffic replicat.				X		X

Table II. Functionality needed on MLD proxy with multiple upstream interfaces per application scenario in mobile networks

## 5 Functional specification of an MLD proxy with multiple interfaces

<To be completed>.

## 6 Security Considerations

<To be completed>.

## 7 IANA Considerations

<IANA considerations text>.

## 8 Conclusions

<To be completed>.

## 9 Acknowledgements

The authors thank Stig Venaas for his valuable comments and suggestions.

The research of Carlos J. Bernardos leading to these results has received funding from the European Community's Seventh Framework Programme (FP7-ICT-2009-5) under grant agreement n. 258053 (MEDIEVAL project), being also partially supported by the Ministry of Science and Innovation (MICINN) of Spain under the QUARTET project (TIN2009-13992-C02-01).

## 10 References

## 10.1 Normative References

- [1] Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "Proxy Mobile IPv6", RFC 5213, August 2008.
- [2] T.C. Schmidt, M. Waehlich, and S. Krishnan, "A Minimal Deployment Option for Multicast Listeners in PMIPv6 Domains", RFC6224, April 2011.
- [3] B. Fenner, H. He, B. Haberman, and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.

## 10.2 Informative References

- [4] J.C. Zuniga, L.M. Contreras, C.J. Bernardos, S. Jeon, Y. Kim, "Multicast Mobility Routing Optimizations for Proxy Mobile IPv6", work in progress, draft-ietf-multimob-pmipv6-ropt-01, September 2012.
- [5] L.M. Contreras, C.J. Bernardos, I. Soto, "PMIPv6 multicast handover optimization by the Subscription Information Acquisition through the LMA (SIAL)", work in progress, draft-ietf-multimob-fast-handover-01, July 2012.
- [6] T.C. Schmidt, M. Waehlich, R. Koodli, G. Fairhurst, "Multicast Listener Extensions for MIPv6 and PMIPv6 Fast Handovers", work in progress, draft-schmidt-multimob-fmipv6-pfmipv6-multicast-06, May 2012
- [7] T.C. Schmidt, S. Gao, H. Zhang, M. Waehlich, "Mobile Multicast Sender Support in Proxy Mobile IPv6 (PMIPv6) Domains", work in progress, draft-ietf-multimob-pmipv6-source-01, July 2012.
- [8] J. Liu, W. Luo, "Routes Optimization for Multicast Sender in Proxy Mobile IPv6 Domain", work in progress, draft-liu-multimob-pmipv6-multicast-ro-02, July 2012.

#### Appendix A. Basic support for multicast listener with PMIPv6

This section briefly summarizes the operation of Proxy Mobile IPv6 [1] and how multicast listener support works with PMIPv6 as specified in [2].

Proxy Mobile IPv6 (PMIPv6) [1] is a network-based mobility management protocol which enables the network to provide mobility support to standard IP terminals residing in the network. These terminals enjoy this mobility service without being required to implement any mobility-specific IP operations. Namely, PMIPv6 is one of the mechanisms adopted by the 3GPP to support the mobility management of non-3GPP terminals in future Evolved Packet System (EPS) networks.

PMIPv6 allows a Media Access Gateway (MAG) to establish a distinct bi-directional tunnel with different Local Mobility Anchors (LMAs), being each tunnel shared by the attached Mobile Nodes (MNs). Each mobile node is associated with a corresponding LMA, which keeps track of its current location, that is, the MAG where the mobile node is attached. IP-in-IP encapsulation is used within the tunnel to forward traffic between the LMA and the MAG. Figure 4 (taken from [1]) shows the architecture of a PMIPv6 domain.

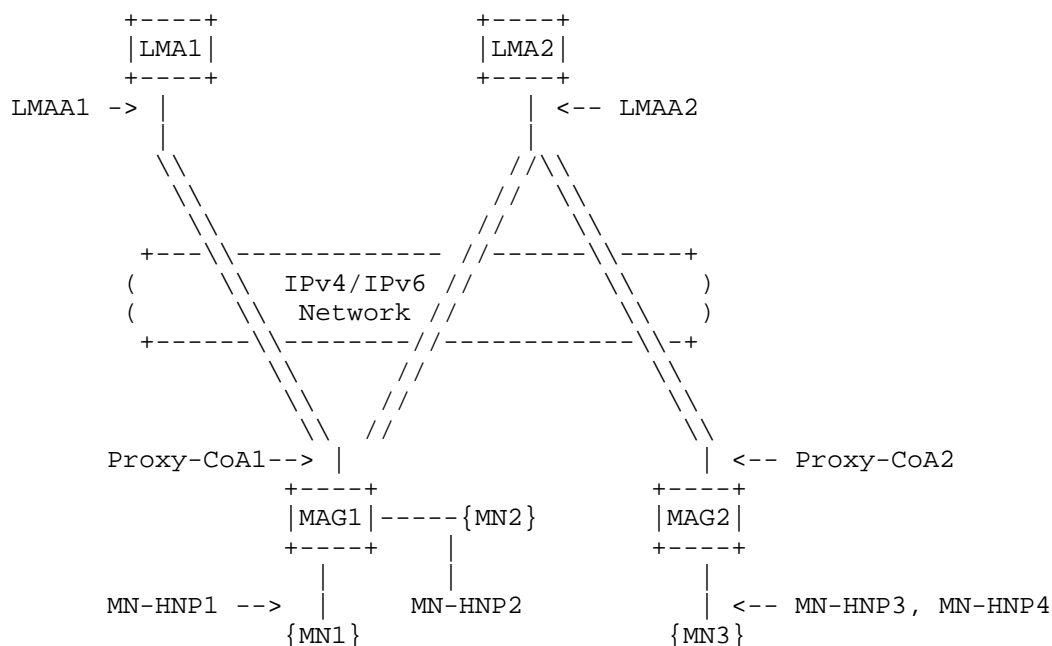


Figure 4. Proxy Mobile IPv6 Domain

The basic solution for the distribution of multicast traffic within a PMIPv6 domain [2] makes use of the bi-directional LMA-MAG tunnels. The base solution follows the so-called remote subscription model, in which the subscribed multicast content is delivered from the Home Network. By doing so, an individual copy of every multicast flow is delivered through the tunnel connecting the mobility anchor to any of the access gateways in the domain. In many cases, these individual copies traverse the same routers in the path towards the access gateways, incurring in an inefficient distribution, equivalent to the unicast distribution of the multicast content in the domain.

The reference scenario for multicast deployment in Proxy Mobile IPv6 domains is illustrated in Figure 5 (taken from [2]).

This fact leads to distribution inefficiencies and higher per-bit delivery costs, incurred by the PMIPv6 domain operator offering transport capabilities to the Home Network operator for serving their MNs when attached to the PMIPv6 domain. As long as the remotely subscribed multicast service is not affected, it seems worthy to explore more optimal ways of distributing such content within the PIMIPv6 domain.



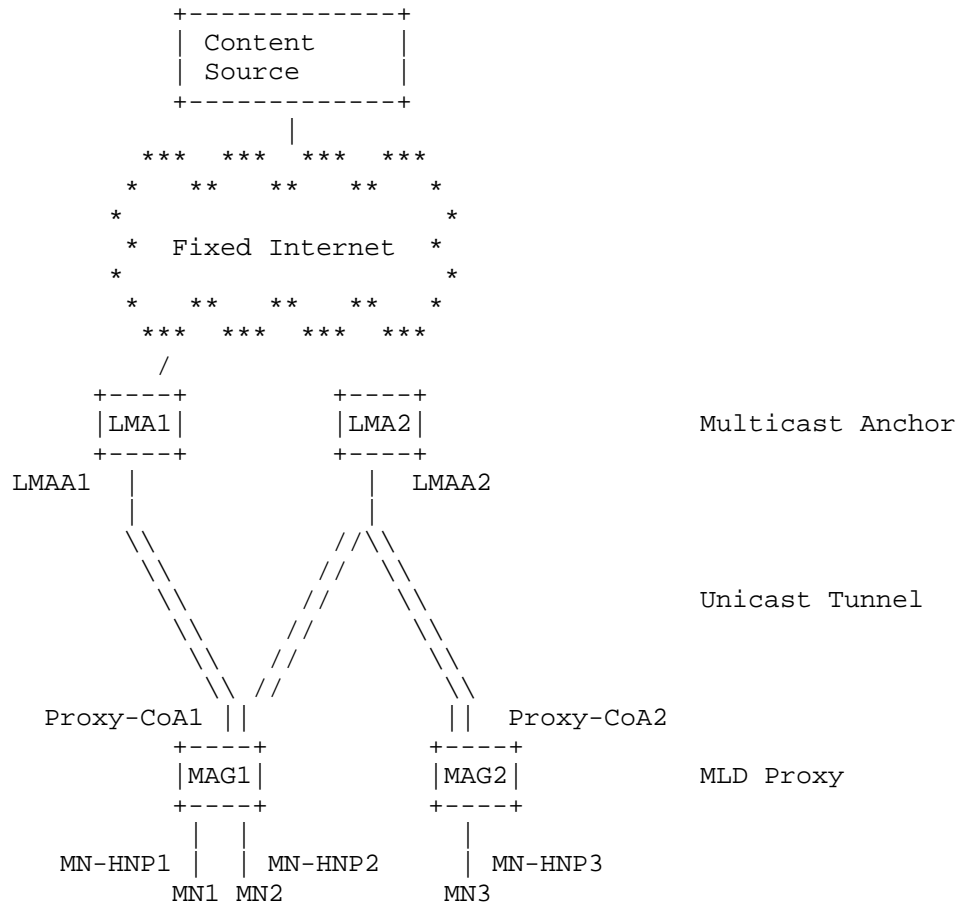


Figure 5. Reference Network for Multicast Deployment in PMIPv6

## Authors' Addresses

Luis M. Contreras  
 Telefonica I+D  
 EMail: lmcm@tid.es

Carlos J. Bernardos  
 Universidad Carlos III de Madrid  
 EMail: cjb@it.uc3m.es

INTERNET DRAFT

MLD proxy with multiple upstream

February 25, 2012

Juan Carlos Zuniga  
InterDigital Communications, LLC  
EMail: JuanCarlos.Zuniga@InterDigital.com

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 29, 2013

Yiqun Cai  
Microsoft  
Sri Vallepalli  
Heidi Ou  
Cisco Systems, Inc.  
Andy Green  
British Telecom  
February 25, 2013

Protocol Independent Multicast DR Load Balancing  
draft-ietf-pim-drlb-02.txt

Abstract

On a multi-access network such as an Ethernet, one of the PIM routers is elected as a Designated Router (DR). The PIM DR has two roles in the PIM protocol. On the first hop network, the PIM DR is responsible for registering an active source to the RP if the group is operated in PIM SM. On the last hop network, the PIM DR is responsible for tracking local multicast listeners and forwarding traffic to these listeners if the group is operated in PIM SM/SSM/DM. In this document, we propose a modification to the PIM protocol that allows more than one of these last hop routers to be selected so that the forwarding load can be distributed to and handled among these routers. A router responsible for forwarding for a particular group is called a Group Designated Router (GDR).

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 29, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the

document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Terminology . . . . .	3
2. Introduction . . . . .	3
3. Applicability . . . . .	6
4. Functional Overview . . . . .	6
4.1. GDR Candidates . . . . .	7
4.2. Hash Mask . . . . .	7
4.3. PIM Hello Options . . . . .	8
5. Packet Format . . . . .	9
5.1. PIM DR Load Balancing Capability (LBC) Hello TLV . . . . .	9
5.2. PIM DR Load Balancing GDR (LBGDR) Hello TLV . . . . .	9
6. Protocol Specification . . . . .	10
6.1. PIM DR Operation . . . . .	10
6.2. PIM GDR Candidate Operation . . . . .	10
6.3. PIM Assert Modification . . . . .	11
7. IANA Considerations . . . . .	12
8. Security Considerations . . . . .	12
9. Acknowledgement . . . . .	12
10. References . . . . .	12
10.1. Normative Reference . . . . .	12
10.2. Informative References . . . . .	13
Authors' Addresses . . . . .	13

## 1. 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 [RFC2119].

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

This document also introduces the following new acronyms:

- o GDR: GDR stands for "Group Designated Router". For each multicast group, a hash algorithm (described below) is used to select one of the routers as GDR. The GDR is responsible for initiating the forwarding tree building for the corresponding group.
- o GDR Candidate: a last hop router that has potential to become a GDR. A GDR Candidate must have the same DR priority as the DR router. It must send and process received new PIM Hello Options as defined in this document. There might be more than one GDR Candidate on a LAN. But only one can become GDR for a specific multicast group.

## 2. Introduction

On a multi-access network such as an Ethernet, one of the PIM routers is elected as a Designated Router (DR). The PIM DR has two roles in the PIM protocol. On the first hop network, the PIM DR is responsible for registering an active source with the RP if the group is operated in PIM SM. On the last hop network, the PIM DR is responsible for tracking local multicast listeners and forwarding to these listeners if the group is operated in PIM SM/SSM/DM.

Consider the following last hop network in Figure 1:

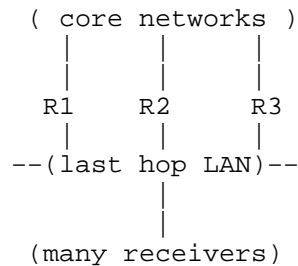


Figure 1: Last Hop Network

Assume R1 is elected as the Designated Router. According to [RFC4601], R1 will be responsible for forwarding to the last hop LAN. In addition to keeping track of IGMP and MLD membership reports, R1 is also responsible for initiating the creation of source and/or shared trees towards the senders or the RPs.

Forcing sole data plane forwarding responsibility on the PIM DR proves a limitation in the protocol. In comparison, even though an OSPF DR, or an IS-IS DIS, handles additional duties while running the OSPF or IS-IS protocols, they are not required to be solely responsible for forwarding packets for the network. On the other hand, on a last hop LAN, only the PIM DR is asked to forward packets while the other routers handle only control traffic (and perhaps drop packets due to RPF failures). The forwarding load of a last hop LAN is concentrated on a single router.

This leads to several issues. One of the issues is that the aggregated bandwidth will be limited to what R1 can handle towards this particular interface. These days, it is very common that the last hop LAN usually consists of switches that run IGMP/MLD or PIM snooping. This allows the forwarding of multicast packets to be restricted only to segments leading to receivers who have indicated their interest in multicast groups using either IGMP or MLD. The emergence of the switched Ethernet allows the aggregated bandwidth to exceed, some times by a large number, that of a single link. For example, let us modify Figure 1 and introduce an Ethernet switch in Figure 2.

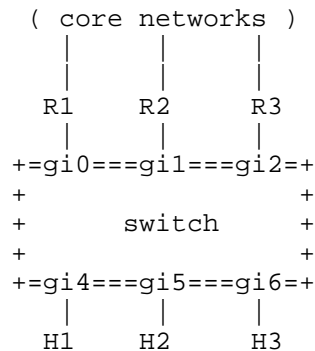


Figure 2: Last Hop Network with Ethernet Switch

Let us assume that each individual link is a Gigabit Ethernet. Each router, R1, R2 and R3, and the switch have enough forwarding capacity to handle hundreds of Gigabits of data.

Let us further assume that each of the hosts requests 500 mbps of data and different traffic is requested by each host. This represents a total 1.5 gbps of data, which is under what each switch or the combined uplink bandwidth across the routers can handle, even under failure of a single router.

On the other hand, the link between R1 and switch, via port gi0, can only handle a throughput of 1gbps. And if R1 is the only router, the PIM DR elected using the procedure defined by RFC 4601, at least 500 mbps worth of data will be lost because the only link that can be used to draw the traffic from the routers to the switch is via gi0. In other words, the entire network's throughput is limited by the single connection between the PIM DR and the switch (or the last hop LAN as in Figure 1).

The problem may also manifest itself in a different way. For example, R1 happens to forward 500 mbps worth of unicast data to H1, and at the same time, H2 and H3 each requests 300 mbps of different multicast data. Once again packet drop happens on R1 while in the mean time, there is sufficient forwarding capacity left on R2 and R3 and link capacity between the switch and R2/R3.

Another important issue is related to failover. If R1 is the only forwarder on the last hop network, in the event of a failure when R1 goes out of service, multicast forwarding for the entire network has to be rebuilt by the newly elected PIM DR. However, if there was a way that allowed multiple routers to forward to the network for different groups, failure of one of the routers would only lead to

disruption to a subset of the flows, therefore improving the overall resilience of the network.

In this document, we propose a modification to the PIM protocol that allows more than one of these routers, called Group Designated Router (GDR) to be selected so that the forwarding load can be distributed to and handled by a number of routers.

### 3. Applicability

The proposed change described in this specification applies to PIM last hop routers only.

It does not alter the behavior of a PIM DR on the first hop network. This is because the source tree is built using the IP address of the sender, not the IP address of the PIM DR that sends the registers towards the RP. The load balancing between first hop routers can be achieved naturally if an IGP provides equal cost multiple paths (which it usually does in practice). And distributing the load to do registering does not justify the additional complexity required to support it.

### 4. Functional Overview

In the existing PIM DR election, when multiple last hop routers are connected to a multi-access network (for example, an Ethernet), one of them is selected to act as PIM DR. The PIM DR is responsible for sending Join/Prune messages to the RP or source. To elect the PIM DR, each PIM router on the network examines the received PIM Hello messages and compares its DR priority and IP address with those of its neighbors. The router with the highest DR priority is the PIM DR. If there are many such routers, their IP addresses are used as the tie breaker, as described in [RFC4601].

In order to share forwarding load among last hop routers, besides the normal PIM DR election, the GDR is also elected on the last hop multi-access network. There is only one PIM DR on the multi-access network, but there might be multiple GDR Candidates.

For each multicast group, a hash algorithm is used to select one of the routers to be the GDR. Hash Masks are defined for Source, Group and RP separately, in order to handle different PIM modes. The masks are announced in PIM Hello by DR as a Load Balancing GDR TLV (LBGDR TLV). Besides that, a Load Balancing Capability TLV (LBC TLV) is also announced by routers support this specification. Last hop routers who are with the new LBC TLV and with the same DR priority as



the PIM DR are GDR Candidates.

A hash algorithm based on the announced Source, Group or RP masks allows one GDR to be assigned to a corresponding multicast group, and that GDR is responsible for initiating the creation of the multicast forwarding tree for the group.

#### 4.1. GDR Candidates

GDR is the new concept introduced by this specification. To become a candidate GDR, a router MUST support this specification and also have the same DR priority as the DR. For example, assume there are 4 routers on the LAN: R1, R2, R3 and R4, which all support this specification. R1, R2 and R3 have the same DR priority while R4's DR priority is less preferred. In this example, only R1, R2 and R3 will be eligible for GDR election. R4 is not because R4 will not become a PIM DR unless all of R1, R2 and R3 go out of service.

Further assume router R1 wins the PIM DR election. In its Hello packet, R1 will include the identity of R1, R2 and R3 (the GDR Candidates) besides its own Load Balancing Hash Masks.

#### 4.2. Hash Mask

A Hash Mask is used to extract a number of bits from the corresponding IP address field (32 for v4, 128 for v6), and calculate a hash value. A hash value is used to select GDR from GDR Candidates advertised by PIM DR. For example, 0.255.0.0 defines a Hash Mask for an IPv4 address that masks the first, the third and the fourth octets.

There are three Hash Masks defined,

- o RP Hash Mask
- o Source Hash Mask
- o Group Hash Mask

The Hash Masks must be configured on the PIM routers that can potentially become a PIM DR.

The hash function used by BSR seems to serve GDR selection well. We use it for now with some modification, and will do more experiments.

For ASM groups, a hash value is calculated using the following BSR style formula:

- o  $\text{hashvalue\_RP}(\text{RP\_address}, \text{RP\_hashmask}, \text{GDR}(i)) = (1103515245 * ((1103515245 * (\text{RP\_address} \& \text{RP\_hashmask}) + 12345) \text{ XOR } \text{GDR}(i)) + 12345) \bmod 2^{31}$

RP\_address is the address of the RP defined for the group. GDR(i) is the address of GDR Candidate.

Similar to BSR hash function, for address families other than IPv4, a 32-bit digest to be used. Such a digest method must be used consistently throughout all GDR Candidates.

If RP\_hashmask is 0, a hash value is also calculated using the group Hash Mask in a similar fashion.

- o  $\text{hashvalue\_G}(\text{Group\_address}, \text{Group\_hashmask}, \text{GDR}(i)) = (1103515245 * ((1103515245 * (\text{Group\_address} \& \text{Group\_hashmask}) + 12345) \text{ XOR } \text{GDR}(i)) + 12345) \bmod 2^{31}$

For SSM groups, a hash value is calculated using both the source and group Hash Mask

- o  $\text{hashvalue\_SG}(\text{Group\_address}, \text{Group\_hashmask}, \text{Source\_address}, \text{Source\_hashmask}, \text{GDR}(i)) = (1103515245 * ((1103515245 * (\text{Group\_address} \& \text{Group\_hashmask}) + 12345) \text{ XOR } (\text{Source\_address} \& \text{Source\_hashmask}) + 12345) \text{ XOR } \text{GDR}(i)) + 12345) \bmod 2^{31}$

The GDR Candidate with the highest hash value is chosen as the GDR. If more than one GDR Candidate has the same highest hash value, the GDR Candidate with the highest address is chosen.

#### 4.3. PIM Hello Options

When a non-DR PIM router that supports this specification sends a PIM Hello, it includes a new option, called "Load Balancing Capability TLV (LBC TLV)".

Besides this new LBC TLV, the elected PIM DR router also includes a "Load Balancing GDR TLV (LBGDR TLV)" in its PIM Hello. The LBGDR TLV consists of three Hash Masks as defined above and the addresses of all GDR Candidates on the last hop network.

The elected PIM DR router uses LBC TLV advertised by all routers on the last hop network to compose its LBGDR TLV. The GDR Candidates use LBGDR TLV advertised by PIM DR router to calculate hash value.

## 5. Packet Format

### 5.1. PIM DR Load Balancing Capability (LBC) Hello TLV

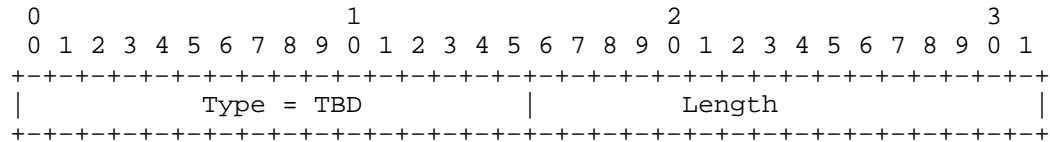


Figure 3: Capability Hello TLV

Type: TBD.  
Length: is zero

This LBC TLV SHOULD be advertised by last hop routers that support this specification.

### 5.2. PIM DR Load Balancing GDR (LBGDR) Hello TLV

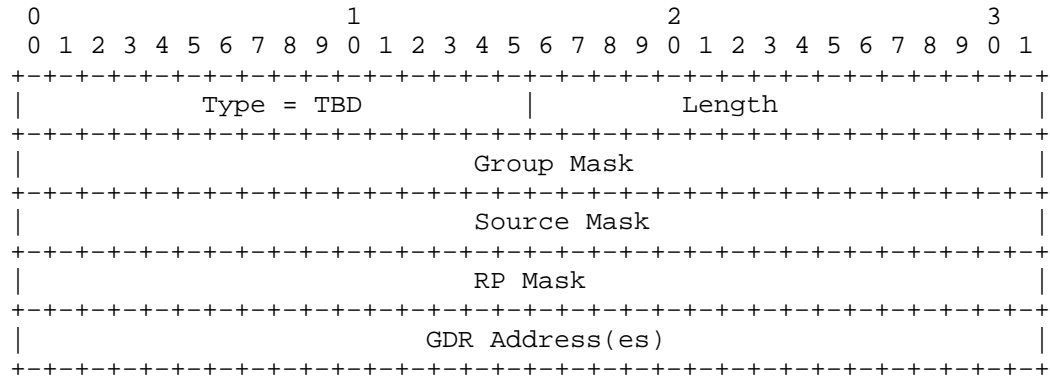


Figure 4: GDR Hello TLV

Type: TBD  
Length:  
Group Mask (32/128 bits): Mask  
Source Mask (32/128 bits): Mask  
RP Mask (32/128 bits): Mask  
All masks MUST be in the same address family, with the same length.  
GDR Address (32/128 bits): Address(es) of GDR Candidates. All addresses must be in the same address family. The addresses are used in hash value calculation.

This LBGDR TLV SHOULD only be advertised by the elected PIM DR router.

## 6. Protocol Specification

### 6.1. PIM DR Operation

LBC TLV indicates the router's capability to support this specification. LBGRD TLV on PIM DR contains value of masks from user configuration, followed by the addresses of all GDR Candidates.

The DR election process is still the same as defined in [RFC4601]. A DR that supports this specification advertises a new Hello Option LBGRD TLV to include all GDR Candidates. Moreover, same as non-DR routers, DR also advertises LBC TLV Hello Option to indicate its capability of supporting this specification.

If a PIM DR receives a neighbor Hello with LBGRD TLV, the PIM DR SHOULD ignore the TLV.

If a PIM DR receives a neighbor Hello with LBC TLV, and the neighbor has the same DR priority as PIM DR itself, the PIM DR SHOULD consider the neighbor as a GDR Candidate and insert the neighbor's address into the list of LBGRD TLV.

### 6.2. PIM GDR Candidate Operation

When an IGMP join is received, without this proposal, router R1 (the PIM DR) will handle the join and potentially run into the issues described earlier. Using this proposal, a hash algorithm is used to determine which router is going to be responsible for building forwarding trees on behalf of the host.

The algorithm works as follows, assuming the router in question is X and a GDR Candidate:

- o If the group is ASM, and if the RP Hash Mask announced by the PIM DR is not 0, calculate the value of hashvalue\_RP. If X results in the highest hashvalue\_RP, X becomes the GDR.
- o If the group is ASM and if the RP Hash Mask announced by the PIM DR is 0, obtain the value of hashvalue\_Group, to decide whether X is the GDR.
- o If the group is SSM, then use hashvalue\_SG to determine if X is the GDR.

If X is the GDR for the group, X will be responsible for building the forwarding tree.

A router that supports this specification advertises LBC TLV in its Hello, even if the router may not be a GDR Candidate.

A GDR Candidate may receive a LBGDR TLV from PIM DR router, with different Hash Masks from those configured on it, The GDR Candidate must use the Hash Masks advertised by the PIM DR Hello to calculate the hash value.

A GDR Candidate may receive an LBGDR TLV from a non-DR PIM router. The GDR candidate must ignore such LBGDR TLV.

A GDR Candidate may receive a Hello from the elected PIM DR, and the PIM DR does not support this specification. The GDR election described by this specification will not take place, that is only the PIM DR joins the multicast tree.

### 6.3. PIM Assert Modification

When routers restart, GDR may change for a specific group, which might cause packet drops.

For example, assume that there are two streams G1 and G2, and R1 is the GDR for G1 and R2 is the GDR for G2. When R3 comes up online, it is possible that R3 becomes GDR for G1 and G2, and rebuilding of the forwarding trees for G1 and G2 will lead to potential packet loss.

This is not a typical deployment scenario but it still might happen. Here we describe a mechanism to minimize the impact.

When the role of GDR changes as above, instead of immediately stopping forwarding, R1 and R2 continue forwarding to G1 and G2 respectively, while in the same time, R3 build forwarding trees for G1 and G2. This will lead to PIM Asserts.

The same tie breakers are used to select an Assert winner with one modification. That is, instead of comparing IP addresses as the last

resort, a router considers whether the sender of an Assert is a GDR. In this example, R1 will let R3 be the assert winner for G1, and R2 will do the same for R3 for G2. This will cause some duplicates in the network while minimizing packet loss.

If a router on the LAN does not support this specification, the Assert modification described above will not take place, that is only the IP address of an Assert sender is used as the tie breaker. For example, if R4, with preferred IP address, does not understand GDR and sends Assert for G1 to R3, which is the GDR for G1, R3 will grant R4 as the Assert winner, and clear OIF on R3.

## 7. IANA Considerations

Two new PIM Hello Option Types are required to be assigned to the DR Load Balancing messages. According to [HELLO-OPT], this document recommends 33(0x21) as the new "PIM DR Load Balancing Capability Hello Option", and 34(0x22) as the new "PIM DR Load Balancing GDR Hello Option".

## 8. Security Considerations

Security of the PIM DR Load Balancing Hello message is only guaranteed by the security of PIM Hello packet, so the security considerations for PIM Hello packets as described in PIM-SM [RFC4601] apply here.

## 9. Acknowledgement

The authors would like to thank Steve Simlo, Taki Millonis for helping with the original idea, Bill Atwood for review comments.

## 10. References

### 10.1. 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.

## 10.2. 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.
- [HELLO-OPT] IANA, "PIM Hello Options", PIM-HELLO-OPTIONS per RFC4601 <http://www.iana.org/assignments/pim-hello-options>, March 2007.

## Authors' Addresses

Yiqun Cai  
Microsoft  
La Avenida  
Mountain View, CA 94043  
USA

Email: [yiqunc@microsoft.com](mailto:yiqunc@microsoft.com)

Sri Vallepalli  
Cisco Systems, Inc.  
Tasman Drive  
San Jose, CA 95134  
USA

Email: [svallepa@cisco.com](mailto:svallepa@cisco.com)

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

Email: [hou@cisco.com](mailto:hou@cisco.com)

Andy Green  
British Telecom  
Adastral Park  
Ipswich IP5 2RE  
United Kingdom

Email: andy.da.green@bt.com





Multimob Working Group  
Internet-Draft  
Intended status: Informational  
Expires: August 26, 2013

H. Liu  
  
M. McBride  
Huawei Technologies  
H. Asaeda  
NICT  
February 22, 2013

IGMP/MLD Optimizations in Wireless and Mobile Networks  
draft-liu-multimob-igmp-mld-wireless-mobile-03

Abstract

This document proposes a variety of optimization approaches for IGMP and MLD in wireless and mobile networks. It aims to provide useful guidelines to allow efficient multicast communication in these networks using IGMP or MLD protocols.

Requirements Language

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].

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 26, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents  
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Requirements . . . . .	3
2.1. Characteristics of Wireless and Mobile Multicast . . . . .	3
2.2. Wireless Link Model . . . . .	4
2.3. Requirements on IGMP and MLD . . . . .	5
3. IGMP/MLD Optimization for Wireless and Mobile Networks . . . . .	6
3.1. Switching Between Unicast and Multicast Queries . . . . .	6
3.2. General Query Supplemented with Unicast Query . . . . .	6
3.3. Retransmission of Queries . . . . .	7
3.4. General Query Suppression . . . . .	7
3.5. Tuning Response Delay According to Link Type and Status . . . . .	8
3.6. Triggering Reports and Queries Quickly During Handover . . . . .	9
4. Applicability and Interoperability Considerations . . . . .	9
5. IGMP/MLD Suspend and Resume . . . . .	10
5.1. IGMP/MLD Suspend Request . . . . .	10
5.2. IGMP/MLD Resume Request . . . . .	10
5.3. IGMP/MLD Suspend Reception . . . . .	10
5.4. IGMP/MLD Resume Reception . . . . .	11
6. Timers, Counters, and Their Default Values . . . . .	12
6.1. IGMP/MLD Suspend Interval Timer . . . . .	12
7. IANA Considerations . . . . .	12
8. Security Considerations . . . . .	12
9. Acknowledgements . . . . .	12
10. References . . . . .	13
10.1. Normative References . . . . .	13
10.2. Informative References . . . . .	14
Authors' Addresses . . . . .	15

## 1. Introduction

The deployments of various wireless access techniques are being combined with the use of video and other applications which rely upon IP Multicast. Wireless and mobile multicast are attracting increasing interest from content and service providers. Multicast faces challenges with dynamic group membership management being under the constant update of delivery paths introduced by node movement. There is a high probability of loss and congestion due to limited reliability and capacity of wireless links.

Multicast networks are generally constructed by the IGMP and MLD group management protocols (respectively for IPv4 and IPv6 networks) to track valid receivers and by multicast routing protocol building multicast delivery paths. This document focuses only on IGMP and MLD, the protocols used by a host to subscribe to a multicast group and the protocols that are most likely to be exposed to wireless links when supporting terminal mobility. As IGMP and MLD were designed for fixed users on a wired link, they do not necessarily work well for different wireless link types and mobile scenarios. IGMP/MLD should be enhanced to be more applicable in these mobile/wireless environments.

This memo proposes a variety of optimizations for IGMP and MLD, in wireless and mobile networks, to improve network performance, with minimum changes on the protocol behavior and without introducing interoperability issues. These solutions can also be applied in wired networks when efficiency or reliability is required.

For generality, this memo does not put limitations on the type of wireless techniques running below IGMP or MLD. They could be cellular, WiMAX, WiFi and etc, and are modeled as different abstract link models as described in section 2.2. Even though some of them (such as WiFi) have multicast limitations, it is probable that IGMP/MLD is enabled on the wireless terminal and multicast is supported across the network. The mobile IP protocol adopted on the core side, upstream from the access router, could be PMIP, MIPv4, or MIPv6.

## 2. Requirements

### 2.1. Characteristics of Wireless and Mobile Multicast

Several limitations should be considered when supporting IP multicast in wireless and mobile networks, including:

O Limited link bandwidth: wireless links usually have limited bandwidth, and the situation will be made even worse if a high volume

of video multicast data has to be carried. Additionally, the bandwidth available in the upstream and downstream directions may be asymmetrical.

O High loss rate: wireless links usually have packet loss ranging from 1% to 30% according to different links types and conditions. Also when packets have to travel between home and access networks (e.g. through a tunnel), they are prone to loss if the two networks are distant from each other.

O Frequent membership change: in fixed multicast, membership change only happens when a user leaves or joins a group, while in mobile scenario membership may also change when a user changes its location.

O Prone to performance degradation: the possible increased interaction of protocols across layers for mobility management, and the limitation of link capacity, may lead to network performance degradation and even to complete connection loss.

O Increased Leave Latency: the leave latency in mobile multicast might be increased due to user movement, especially if the traffic has to be transmitted between access and home networks, or if there is a handshake between networks.

## 2.2. Wireless Link Model

Wireless links can typically be categorized into three models: point-to-point (PTP), point-to-multipoint (PTMP), and broadcast link models.

In the PTP model, one link is dedicated for two communication facilities. For multicast transmission, each PTP link normally has only one receiver and the bandwidth is dedicated for that receiver. Such link model may be implemented by running PPP on the link or having separate VLAN assignment for each receiver. In a mobile network, a tunnel between entities of home and foreign networks should be recognized as a PTP link.

PTMP is the model for multipoint transmission wherein there is one centralized transmitter and multiple distributed receivers. PTMP provides common downlink channels for all receivers and dedicated uplink channel for each receiver. Bandwidth downstream is shared by all receivers on the same link.

Broadcast links can connect two or more nodes and support broadcast transmissions. It is quite similar to fixed Ethernet link model and its link resource is shared in both uplink and downlink directions.

### 2.3. Requirements on IGMP and MLD

IGMP and MLD are usually run between mobile or wireless terminals and their first-hop access routers (i.e. home or foreign routers) to subscribe to a IP multicast channel. Currently the version in-use includes IGMPv2 [RFC2236] and its IPv6 counterpart MLDv1 [RFC2710], IGMPv3 [RFC3376] and its IPv6 counterpart MLDv2 [RFC3810], and LW-IGMPv3/MLDv2 [RFC5790]. All these versions have basic group management capability required by a multicast subscription. The differences lie in that IGMPv2 and MLDv1 can only join and leave a non-source-specific group, while IGMPv3 and MLDv2 can select including and excluding specific sources for their join and leave operation, and LW-IGMPv3/MLDv2 simplifies IGMPv3/MLDv2 procedures by discarding excluding-source function. Among these versions, (LW-) IGMPv3/MLDv2 has the capability of explicitly tracking each host member.

From the illustration given in section 2.1 and 2.2, it is desirable for IGMP and MLD to have the following characteristics when used in wireless and mobile networks:

- o Adaptive to link conditions: wireless networks have various link types, each with different bandwidth and performance features. IGMP or MLD should be able to be adaptive to different link models and link conditions to optimize its protocol operation.
- o Minimal group join/leave latency: because mobility and handover may cause a user to join and leave a multicast group frequently, fast join and leave by the user helps to accelerate service activation and to release unnecessary resources quickly to optimize resource utilization.
- o Robust to packet loss: the unreliable packet transmission due to instable wireless link conditions and limited bandwidth, or long distance transmission in mobile network put more strict robustness requirement on delivery of IGMP and MLD protocol messages.
- o Reducing packet exchange: wireless link resources are usually more limited, precious, and congested compared to their wired counterpart. This requires packet exchange be minimized without degrading protocol performance.
- o Packet burst avoidance: large number of packets generated in a short time interval may have the tendency to deteriorate wireless network conditions. IGMP and MLD should be optimized to reduce the probability of packet burst.

### 3. IGMP/MLD Optimization for Wireless and Mobile Networks

This section introduces several optimizations for IGMP and MLD in wireless or mobile environment. The aim is to meet the requirements described in section 2.3. It should be noted that because an enhancement in one direction might result in weakening effect in another, balances should be taken cautiously to realize overall performance elevation.

#### 3.1. Switching Between Unicast and Multicast Queries

IGMP/MLD protocols use multicast Queries whose destinations are multicast addresses and also allows use of unicast Query with unicast destination to be sent only to one host. Unicast Query has the advantage of not affecting other hosts on the same link, and is desirable for wireless communication because a mobile terminal often has limited battery power [RFC6636]. But if the number of valid receivers is large, using unicast Query for each receiver is inefficient because large number of Unicast Queries have to be generated, in which situation normal multicast Query will be a good choice because only one General Query is needed. If the number of receivers to be queried is small, unicast Query is advantageous over the multicast one.

More flexibly, the router can choose to switch between unicast and multicast Queries according to the practical network conditions. For example, if the receiver number is small, the router could send unicast Queries respectively to each receiver, without arousing other non-member terminal which is in dormant state. When the receiver number reaches a predefined level, the router could change to use multicast Queries. To have the knowledge of the number of the valid receivers, a router is required to enable explicit tracking, and because Group-Specific Query and Group-and-Source-Specific Query are usually not used under explicit tracking [RFC6636], the switching operation mostly applies to General Queries.

#### 3.2. General Query Supplemented with Unicast Query

The Unicast Query can be used in assistance to General Query to improve the robustness of solicited reports when General Query fails to collect all of its valid members. It requires the explicit tracking to be enabled and can be used when a router after sending a periodical General Query collects successfully most of the valid members' responses while losing some of which are still valid in its database. This may be because these reports are not generated or generated but lost for some unknown reasons. The router could choose to unicast a Query respectively to each non-respondent valid receiver to check whether they are still alive for the multicast reception,

without affecting the majority of receivers that have already responded. Unicast Queries under this condition could be sent at the end of the [Maximum Response Delay] after posting a General Query, and be retransmitted for [Last Member Query Count] times, at an interval of [Last Member Query Interval].

### 3.3. Retransmission of Queries

In IGMP and MLD, apart from the continuously periodical transmission, General Query is also transmitted during a router's startup. It is transmitted for [Startup Query Count] times by [Startup Query Interval]. There are some other cases where retransmission of General Query is beneficial which are not covered by current IGMP and MLD protocols as shown as following.

For example, a router which keeps track of all its active receivers, if after sending a General Query, fails to get any response from the receivers which are still valid in its membership database. This may be because all the responses of the receivers happen to be lost, or the sent Query does not arrive at the other side of the link to the receivers. The router could compensate this situation by retransmitting the General Query to solicit its active members. The retransmission can also be applied to Group-Specific or Group-and-Source-Specific Query on a router without explicit tracking capability, when these Specific Queries cannot collect valid response, to prevent missing valid members caused by lost Queries and Reports.

The above compensating Queries could be sent [Last Member Query Count] times, at the interval of [Last Member Query Interval], if the router cannot get any feedback from the receivers.

### 3.4. General Query Suppression

In IGMP and MLD, the General Query is sent periodically and continuously without any limitation. It helps soliciting the state of current valid member but has to be processed by all hosts on the link, whether they are valid multicast receivers or not. When there is no receiver, the transmission of the General Query is a waste of resources for both the host and the router.

An IGMP/MLD router could suppress its transmission of General Query if it knows there is no valid multicast receiver on an interface, e.g. in the following cases:

O When the last member reports its leave for a group. This could be judged by an explicit tracking router checking its membership database, or by a non-explicit-tracking router getting no response



after sending Group-Specific or Group-and-Source-Specific Query.

O When the only member on a PTP link reports its leaving

O When a router after retransmitting General Queries on startup fails to get any response

O When a router previously has valid members but fails to get any response after several rounds of General Queries.

In these cases the router could make the decision that no member is on the interface and totally stop its transmission of periodical General Queries. If afterwards any valid member joins a group, the router could resume the original cycle of general Querying. Because General Query influences all hosts on a link, suppressing it when it is not needed is beneficial for both the link efficiency and terminal power saving.

### 3.5. Tuning Response Delay According to Link Type and Status

IGMP and MLD use delayed response to spread unsolicited Reports from different hosts to reduce possibility of packet burst. This is implemented by a host responding to a Query in a specific time randomly chosen between 0 and [Maximum Response Delay], the latter of which is determined by the router and is carried in Query messages to inform the hosts for calculation of the response delay. A larger value will lessen the burst better but will increase leave latency (the time taken to cease the traffic flowing after the last member requests the escaping of a channel).

In order to avoid message burst and reduce leave latency, the Response Delay may be dynamically calculated based on the expected number of responders, and link type and status, as shown in the following:

O If the expected number of reporters is large and link condition is bad, longer Maximum Response Delay is recommended; if the expected number of reporters is small and the link condition is good, smaller Maximum response Delay should be set.

o If the link type is PTP, the Maximum Response Delay can be chosen smaller, whereas if the link is PTMP or broadcast medium, the Maximum Response Delay can be configured larger.

The Maximum Response Delay could be configured by the administrator as mentioned above, or be calculated automatically by a software tool implemented according to experiential model for different link modes. The measures to determine the instant value of Maximum Response Delay

are out of this document's scope.

### 3.6. Triggering Reports and Queries Quickly During Handover

When a mobile terminal is moving from one network to another, if it is receiving multicast content, its new access network should try to deliver the content to the receiver without disruption or performance deterioration. In order to implement smooth handover between networks, the terminal's membership should be acquired as quickly as possible by the new access network.

An access router could trigger a Query to a terminal as soon as it detects the terminal's attaching on its link. This could be a General Query if the number of the entering terminals is not small (e.g when they are simultaneously in a moving train). Or this Query could also be a unicast Query for this incoming terminal to prevent unnecessary action of other terminals in the switching area.

For the terminal, it could send a report immediately if it is currently in the multicast reception state, when it begins to connect the new network. This helps establishing more quickly the membership state and enable faster multicast stream injection, because with the active report the router does not need to wait for the query period to acquire the terminal's newest state.

## 4. Applicability and Interoperability Considerations

Among the optimizations listed above, 'Switching between unicast and multicast Queries'(3.1) and 'General Query Supplemented with Unicast Query'(3.2) requires a router to know beforehand the valid members connected through an interface, thus require explicit tracking capability. An IGMP/MLD implementation could choose any combination of the methods listed from 3.1 to 3.6 to optimize multicast communication on a specific wireless or mobile network.

For example, an explicit-tracking IGMPv3 router, can switch to unicast General Queries if the number of members on a link is small (3.1), can trigger unicast Query to a previously valid receiver if failing to get expected responses from it (3.2), can retransmit a General Query if after the previous one cannot collect reports from all valid members (3.3), and can stop sending a General Query when the last member leaves the group (3.4), and etc.

For interoperability, it is required if multiple multicast routers are connected to the same network for redundancy, each router are configured with the same optimization policy to synchronize the membership states among the routers.

## 5. IGMP/MLD Suspend and Resume

### 5.1. IGMP/MLD Suspend Request

IGMP/MLD Suspend is an operation triggered by a host that subscribes multicast channels or an IGMP/MLD proxy [refs.Proxy] to which hosts subscribing multicast channels attached. An IGMP/MLD Suspend message requests an adjacent upstream router to suspend forwarding subscribed data, while to keep the subscription state (i.e., not to prune the routing path). It is useful especially in a mobile network. When a mobile host moves from a current network (i.e., a mobile host's point of attachment) to a different network, an IGMP/MLD Suspend message is sent by the host itself (or an IGMP/MLD proxy to which mobile hosts attached).

When an IGMP/MLD proxy receives IGMP/MLD Suspend messages on its downstream interface, it forwards the Suspend message to its upstream router via its upstream interface if needed (see Section 5.3).

### 5.2. IGMP/MLD Resume Request

When a host that has subscribed multicast channels and sent IGMP/MLD Suspend messages attaches to a new network, it immediately sends IGMP/MLD Resume messages to request its upstream router to resume forwarding the data. The Resume Records specified in the IGMP/MLD Resume message will be the same as that of the Suspend Records the host sent.

### 5.3. IGMP/MLD Suspend Reception

When a multicast router receives an IGMP/MLD Suspend message from the downstream member hosts or IGMP/MLD proxy, it examines whether the message sender is the sole member of the reported channels at the downstream link or not. There are two ways to know it. One is done by the Group-Specific or Group-and-Source Specific Queries. The other is done by the explicit tracking function [refs.explicit].

The router sends the Group-Specific or Group-and-Source Specific Queries for all records in the Suspend messages. If the router receives IGMP/MLD reports including some or all of the Suspend Records, it eliminates the reported records from the Suspend Records and keeps forwarding these data. If the router does not receive IGMP/MLD reports for some or all of the Suspend Records, it recognizes that the Suspend message sender is the sole member host for these channels on the link. After the router organizes the new Suspend Records (that eliminate reported records from the original one), it suspends data forwarding for them.

The explicit tracking function gives advantage of organizing the new Suspended Records. If a router enables the explicit tracking function, it can recognize whether the message sender host is the sole member without sending the Group-Specific or Group-and-Source Specific Queries. Then the router suspends data forwarding based on the up-to-date Suspend Records.

The multicast router maintains Suspend Records until it receives the corresponding IGMP/MLD Resume message (described in the next section) or the IGMP/MLD Suspend Interval timer (described in Section 6.1) is expired. When either the Resume message reception or the timer expiration occurs, the router resume data forwarding for the Suspend Records and discards the Suspend Records.

If a multicast router receives an IGMP/MLD Suspend message, which includes Multicast Address Records already suspended, the router restarts the IGMP/MLD Suspend Interval timer for the corresponding Multicast Address Records.

When an IGMP/MLD proxy receives an IGMP/MLD Suspend message from a downstream host, it behaves as a multicast router as described above, because the proxy device performs the router portion of the IGMP or MLD protocol on its downstream interfaces.

When a mobile node that has sent the IGMP/MLD Suspend message receives the corresponding Group-Specific or Group-and-Source Specific Queries for the Suspend Records, it replies the standard IGMP/MLD Report messages as defined in [refs.IGMPv3][refs.MLDv2].

#### 5.4. IGMP/MLD Resume Reception

When a multicast router receives an IGMP/MLD Resume message, the router examines the message sender and an IGMP/MLD Suspend Interval timer. If the router has the Suspend Records given from the Resume message sender, it compares the Suspend Records with the notified Multicast Address Records specified in the Resume message. For the matched Multicast Address Records, the router then removes the entries from the Suspend Records and resumes data forwarding with restarting the group or source timers. For the mismatched Multicast Address Records, the router keeps unchanged (then will be removed by timeout) or explicitly starts the leave (or prune) procedures for the channels, while it depends on the implementation.

If either the router does not have the corresponding Suspend Records or the IGMP/MLD Suspend Interval timer has expired, then the router does not take any action.

If a router that did not recognize an IGMP/MLD Suspend message (e.g.,

due to packet loss or some troubles in its transmission) receives an IGMP/MLD Resume message, it will accept the message as a regular Current-State Report IGMP/MLD message.

## 6. Timers, Counters, and Their Default Values

### 6.1. IGMP/MLD Suspend Interval Timer

After a multicast router receiving an IGMP/MLD Suspend message will identify the corresponding multicast sessions/channels, it suspends data forwarding and keeps the Suspend Records until the given amount of timer value is expired. This timer is named the "IGMP/MLD Suspend Interval timer", which is a configurable value.

The Suspend Interval is used to allow a multicast router to resume the multicast session. Therefore, if the multicast router does not receive the corresponding IGMP/MLD Resume message for the IGMP/MLD Resume operation within the Suspend Interval, it leaves the sessions/channels recorded in the Suspend Records and discards the Suspend Records. Note that the router does not send any IGMP/MLD Query message for the timeout sessions/channels and immediately leaves from them.

## 7. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

## 8. Security Considerations

Since the methods only involve the tuning of protocol behavior by e.g. retransmission, changing delay parameter, or other compensating operations, they do not introduce additional security weaknesses. The security considerations described in [RFC2236], [RFC3376], [RFC2710] and [RFC3810] can be reused. And to achieve some security level in insecure wireless network, it is possible to take stronger security procedures during IGMP/MLD message exchange, which are out of the scope of this memo.

## 9. Acknowledgements

The authors would like to thank Behcet Sarikaya, Qin Wu, Stig Venaas,

Gorry Fairhurst, Thomas C. Schmidt, Marshall Eubanks, Suresh Krishnan, J. William Atwood, WeeSan Lee, Imed Romdhani, Liu Yisong and Wei Yong for their valuable comments and suggestions on this document.

## 10. References

### 10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2236, November 1997.
- [RFC2710] Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, October 1999.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [RFC5790] 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.
- [RFC6636] Asaeda, H., Liu, H., and Q. Wu, "Tuning the Behavior of the Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) for Routers in Mobile and Wireless Networks", RFC 6636, May 2012.
- [refs.IGMPv1] Deering, S., "Host Extensions for IP Multicasting", RFC 1112, August 1989.
- [refs.IGMPv2] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2373, July 1997.
- [refs.IGMPv3] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version

3", RFC 3376, October 2002.

[refs.KEYWORDS]

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

[refs.LW] 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.

[refs.MLDv1]

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

[refs.MLDv2]

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

[refs.PIM]

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

[refs.Proxy]

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.

## 10.2. Informative References

[refs.MIPv6]

Johnson, D., Perkins, C., and J. Arkko, "Mobility Support in IPv6", RFC 3775, June 2004.

[refs.Noel]

Jelger, C. and T. Noel, "Multicast for Mobile Hosts in IP Networks: Progress and Challenges", IEEE Wireless Comm. pp.58-64, October 2002.

[refs.PMIPv6]

Gundavelli, S, Ed., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "Proxy Mobile IPv6", RFC 5213, August 2008.

[refs.explicit]

Asaeda, H., "IGMP/MLD-Based Explicit Membership Tracking Function for Multicast Routers",  
draft-ietf-pim-explicit-tracking-04.txt (work in progress), January 2013.

Authors' Addresses

Hui Liu

Email: liu\_helen@126.com

Mike McBride  
Huawei Technologies  
2330 Central Expressway  
Santa Clara CA 95050  
USA

Email: michael.mcbride@huawei.com

Hitoshi Asaeda  
National Institute of Information and Communications Technology (NICT)  
Network Architecture Laboratory  
4-2-1 Nukui-Kitamachi  
Koganei, Tokyo 184-8795  
Japan

Email: asaeda@nict.go.jp



Network Working Group  
Internet-Draft  
Updates: 5384 (if approved)  
Intended status: Standards Track  
Expires: August 10, 2013

S. Venaas  
I. Kouvelas  
J. Arango  
Cisco Systems  
February 6, 2013

Hierarchical Join/Prune Attributes  
draft-venaas-pim-hierarchicaljoinattr-00.txt

Abstract

This document defines a hierarchical method of encoding Join attributes, providing a more efficient encoding when the same attribute values need to be specified for multiple sources in a PIM Join/Prune message.

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 10, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Requirements Notation . . . . .	4
3. Hierarchical Join/Prune Attribute Definition . . . . .	5
4. PIM Address Encoding Types . . . . .	8
5. Hierarchical Join/Prune Attribute Hello Option . . . . .	9
6. Security Considerations . . . . .	10
7. IANA Considerations . . . . .	11
8. Acknowledgments . . . . .	12
9. Normative References . . . . .	13
Authors' Addresses . . . . .	14

## 1. Introduction

PIM Join attributes as defined in [RFC5384] allow for specifying a set of attributes for each of the joined or pruned sources in a PIM Join/Prune message. Attributes must be separately specified for each individual source in the message. However, in some cases the same attributes and values need to be specified for some, or even all, the sources in the message. The attributes and their values then need to be repeated for each of the sources where they apply.

This document provides a hierarchical way of encoding attributes and their values in a Join/Prune message, so that if the same attribute and value is to apply for all the sources, it needs only be specified once in the message. Similarly, if all the sources in a specific group set share a specific attribute and value, it needs only be specified once for the entire group set.

This document updates [RFC5384] which defines an encoding to be used for Encoded-Source Addresses. This document extends this by specifying the same encoding type also for Encoded-Unicast and Encoded-Group formats. This document defines a new IANA registry for PIM encoding types which is to be used for all the fields in PIM messages where encoding types are used, replacing the old registry that is specific to Encoded-Source Addresses. The encoding type used for Join attributes is however still limited to be used in Join/Prune messages. Note that Join attributes, as they are referred to in [RFC5384], also apply to pruned sources in a Join/Prune message. Thus the more correct name Join/Prune attributes will be used throughout the rest of this document.

This document allows Join/Prune attributes to be specified in the Upstream Neighbor Address field, and also in the Multicast Group Address field, of a Join/Prune message. It defines how this is used to specify the same Join/Prune attribute and value for multiple sources. This document also introduces a new Hello Option to indicate support for the hierarchical encoding specified.

## 2. 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].

### 3. Hierarchical Join/Prune Attribute Definition

The format of a PIM Join/Prune message is defined in [RFC4601] as follows:

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								
PIM Ver										Type										Reserved										Checksum									
Upstream Neighbor Address (Encoded-Unicast format)																																							
Reserved										Num groups										Holdtime																			
Multicast Group Address 1 (Encoded-Group format)																																							
Number of Joined Sources																				Number of Pruned Sources																			
Joined Source Address 1 (Encoded-Source format)																																							
:																																							
Joined Source Address n (Encoded-Source format)																																							
Pruned Source Address 1 (Encoded-Source format)																																							
:																																							
Pruned Source Address n (Encoded-Source format)																																							
:																																							
Multicast Group Address m (Encoded-Group format)																																							
Number of Joined Sources																				Number of Pruned Sources																			
Joined Source Address 1 (Encoded-Source format)																																							
:																																							
Joined Source Address n (Encoded-Source format)																																							
Pruned Source Address 1 (Encoded-Source format)																																							
:																																							
Pruned Source Address n (Encoded-Source format)																																							

The message contains a single Upstream Neighbor Address, and one or more group sets. Each group set contains a Group Address and two source lists, the Joined Sources and the Pruned Sources. The Upstream Neighbor Address, the group addresses and the source addresses are all encoded in Encoded-Unicast format, Encoded-Group format and Encoded-Source format, respectively. In this document we make use of this to allow Join/Prune attributes in each of these addresses, using the encoding in Section 4.

For a Join/Prune message we define a hierarchy of Join/Prune attributes. At the highest level, that is the least specific, we have attributes that apply to every source in the message. These are encoded in the Upstream Neighbor Address. At the next more specific level we have attributes that apply to every source in a group set. They are encoded in a Group Address. And finally at the most specific level, we have attributes that just apply to a single source, encoded in the source address as defined in [RFC5384].

The complete set of attributes that apply to a given source is obtained by combining the message wide attributes, the attributes of the group set that the source belongs to, and the source specific attributes. However, if the same attribute is specified at multiple levels, then the one at the most specific level overrides the other instances of the attribute.

Note that Join/Prune attributes are still applied to sources as specified in [RFC5384]. This document does not change the meaning of any attributes, it is simply a more compact way of encoding an attribute when the same attribute and value applies to multiple sources.

#### 4. PIM Address Encoding Types

Addresses in PIM messages are specified together with an address family and an encoding type. This applies to Encoded-Unicast, Encoded-Group and Encoded-Source addresses. The encoding types allow the address to be encoded according to different schemes. While it is possible to have the same encoding type value indicate different encodings depending on whether it is a Unicast, Group or Source address, it is simpler to have the same encoding type value indicate the same encoding independent of where it is used. This means that as currently defined, 0 means a native encoding, and 1 means there are Join/Prune attributes, encoded according to [RFC5384]. Even if the encoding type space is shared between the different address types (Encoded-Unicast, Encoded-Group and Encoded-Source), one could have a specific encoding apply to a specific address type if needed.

The current IANA PIM Encoded-Source Address Encoding Type Field registry should be changed into a PIM Address Encoding Type registry.



## 5. Hierarchical Join/Prune Attribute Hello Option

A PIM router indicates that it supports the mechanism specified in this document by including the Hierarchical Join/Prune Attribute Hello Option in its PIM Hello message. Note that it also needs to include the Join-Attribute Hello option as specified in [RFC5384]. The format of the Hierarchical Join/Prune Attribute Hello Option is defined to be:

```

      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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|           OptionType = TBD           |           OptionLength = 0           |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

OptionType = TBD, OptionLength = 0. Note that there is no option value included.

A PIM router MUST NOT send a Join/Prune message with Join/Prune attributes encoded in the Upstream Neighbor Address or any of the group addresses out any interface on which there is a PIM neighbor that has not included this option in its Hellos. Even a router that is not the upstream neighbor must be able to parse the message in order to do Join suppression or Prune overriding.

## 6. Security Considerations

This document specifies a more compact encoding of Join/Prune attributes. Use of the encoding has no impact on security.

## 7. IANA Considerations

The current PIM Encoded-Source Address Encoding Type Field registry should be changed into a PIM Address Encoding Type registry. The only required change is the name of the registry. The contents remain the same.

A new PIM Hello Option type needs to be assigned. The string TBD needs to be replaced with the permanently assigned value.

## 8. Acknowledgments

Acknowledgments to be added.

## 9. 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.
- [RFC5384] Boers, A., Wijnands, I., and E. Rosen, "The Protocol Independent Multicast (PIM) Join Attribute Format", RFC 5384, November 2008.

Authors' Addresses

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

Email: stig@cisco.com

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

Email: kouvelas@cisco.com

Jesus Arango  
Cisco Systems  
Tasman Drive  
San Jose, CA 95134  
USA

Email: jearango@cisco.com



PIM Working Group  
Internet Draft  
Intended status: Standards Track  
Expires: September 10, 2013

Hong-Ke Zhang  
Shuai Gao  
Beijing Jiaotong University  
T C.Schmidt  
HAW Hamburg  
Bo-hao Feng  
Li-Li Wang  
Beijing Jiaotong University  
March 11, 2013

Multi-Upstream Interfaces IGMP/MLD Proxy  
draft-zhang-pim-muiimp-00.txt

## Abstract

In this document, followed by the idea mentioned in [4] and subsequent update in [5], an IGMP/MLD proxy with multiple upstream interfaces called MUIIMP is proposed and analyzed. The MUIIMP inherits the basic rule of the IGMP/MLD proxy but extends with multiple upstream interfaces. To avoid data redundancy, each upstream interface of an MUIIMP device MUST NOT send or subscribe the same data simultaneously.

## Requirements Language

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].

## Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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



The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on September, 2013.

#### Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Introduction.....	3
2. Terminology.....	3
3. MUIIMP Behavior.....	4
3.1. The selection of default upstream interface.....	5
3.2. Report of downstream subscriptions to upstream interfaces..	5
3.3. Handover of the upstream interface.....	6
4. Security Considerations.....	6
5. References.....	6
Acknowledgment.....	8

## 1. Introduction

RFC 4605 [1] specifies an IGMP/MLD proxy mechanism for forwarding based solely upon IGMP/MLD membership information in scenarios where multicast routing is not available. According to [1], an IGMP/MLD Proxy performs the router portion of the IGMP/MLD protocol on its downstream interfaces, and the host portion of the IGMP/MLD protocol on its single upstream interface.

The IGMP/MLD proxy mechanism can effectively extend the multicast scope and greatly simplify the implementation of edge devices. However, the IGMP/MLD proxy may exhibit inefficiency in some specific scenarios due to the limitation of single upstream interface. For example, in PMIPv6 multicast environment, multiple IGMP/MLD proxy instances need to be deployed at the MAG in [6], which may result in tunnel convergence problem. In addition, there are also requirements to extend the IGMP/MLD proxy to support multiple upstream interfaces as the emergence of multi-homing.

One thing to note is the idea about multiple upstream interfaces for IGMP/MLD proxy was firstly proposed in the draft [4] to improve the performance of mobile multicast source. The Multimob working group draft [5] includes the related latest descriptions. Considering the multiple upstream interfaces extension is not only required for mobile multicast sources scenarios, this document is presented here.

In this document, an IGMP/MLD proxy with multiple upstream interfaces called MUIIMP is proposed and described. The MUIIMP inherits the basic rule of the IGMP/MLD proxy but extends with multiple upstream interfaces. To avoid data redundancy, each upstream interfaces of an MUIIMP device MUST NOT send or subscribe the same data simultaneously. The MUIIMP is designed to support local multicast listeners and senders.

## 2. Terminology

**Upstream Interface:** A proxy device's interface in the direction of the root of the tree.

**Downstream Interface:** Each of a proxy device's interfaces that is not in the direction of the root of the multicast tree.

**Default upstream interface:** An upstream interface which is by default associated with each downstream node subscribing or sending specific channel (group address prefix) or special multicast state.

### 3. MUIIMP Behavior

The MUIIMP inherits the basic rule of the IGMP/MLD proxy but extends with multiple upstream interfaces. A MUIIMP device has one or more upstream as well as downstream interfaces, which may be any type interfaces, including physical or logical interfaces.

The MUIIMP performs the router portion of the IGMP/MLD protocol on its downstream interfaces, and the host portion of IGMP/MLD on its upstream interfaces. The MUIIMP device **MUST NOT** perform the router portion of IGMP/MLD on its upstream interfaces.

The MUIIMP device maintains a database for multicast listeners consisting of the merger of all subscriptions on any downstream interface. In order to avoid the redundant multicast traffic, the proxy device should initiate unique traffic subscriptions. Besides, a policy list that records the default upstream interface for the downstream nodes is held for the selection of upstream interface.

In the following, the MUIIMP device behavior will be discussed according the role of the downstream nodes.

#### 1) Multicast listener on the downstream interface

Multicast listener reports are group-wise aggregated by the MLD proxy. The aggregated report is issued to the upstream interface based on the subscriptions as well as the policy list. When receiving the IGMP/MLD subscriptions on the downstream interface, the MUIIMP checks the membership database to make a decision whether sends IGMP/MLD membership reports on the corresponding default upstream interface or not. Refer to Section 3.2 for the details about membership subscriptions lookup and report decisions.

When receiving packets on its upstream interfaces, the MUIIMP forwards the traffic to all the downstream interfaces based upon the downstream interfaces' subscriptions.

#### 2) Multicast source on the downstream interface

When receiving packets on its downstream interface, the MUIIMP forwards the traffic to the corresponding default upstream interface, as well as all the downstream interfaces other than the incoming interface based upon the downstream interfaces' subscriptions.

The (first) multicast router(s) operating multicast routing protocol like PIM-SM[7] connected to the outside multicast domain should be configured to treat the multicast source inside the MUIIMP domain

being directly connected. Otherwise, it will discard the data due to the failure of the direct connection check.

### 3.1. The selection of default upstream interface

Typically, the choice of the default upstream interface is based on the policy list which is maintained at the MUIIMP.

The expression of the policy list is like below:

(node prefix, multicast group address/multicast state, upstream interface)

Here node prefix represents the address prefix of the node on the downstream interface that may be a multicast listener or multicast source. And the multicast group address indicates the channel that the multicast listener is subscribing or the multicast source is publishing while the multicast state is only valid for listeners indicating the state about both multicast source and multicast group they are subscribing.

In other word, in the MUIIMP, the multicast group address/multicast state and the node prefix will act as rules to select the default upstream interface. Alternate configurations (e.g., the MAG-LMA tunnel interface in PMIPv6 environment) MAY be applied.

### 3.2. Report of downstream subscriptions to upstream interfaces

To avoid the redundant multicast traffic, the proxy device MUST NOT send the same multicast subscription record on different upstream interfaces simultaneously. In detail, we recommend the following rules when receiving an IGMP/MLD subscription on the downstream interface.

- 1) If the received IGMP/MLD subscription is new and has not been subscribed by other downstream multicast listeners, the proxy device SHOULD initiate the IGMP/MLD subscription on the corresponding default upstream interface.
- 2) If there exists the same IGMP/MLD subscription which has already been subscribed by other downstream multicast listener, the proxy device SHOULD not initiate extra IGMP/MLD subscription.
- 3) If there exists IGMP/MLD subscriptions which have already included the received IGMP/MLD subscription, the proxy device SHOULD not initiate extra IGMP/MLD subscription.

- 4) If there exists overlapping subsets between the received IGMP/MLD subscription and current IGMP/MLD subscriptions, the proxy device SHOULD initiate the IGMP/MLD subscription on the corresponding default upstream interface excluding the overlapping subsets that have been subscribed before.

All subscriptions sent on the same upstream interface SHOULD be merged according the merging rule in RFC 4605. In addition, the local multicast source should be excluded in the final subscriptions to avoid replicated multicast traffic from outside.

### 3.3. Handover of the upstream interface

If an upstream interface fails for some reason such as the deletion of the tunnel interface in mobile environment, the handover of the upstream interface is performed. Generally, all the subscriptions sent on the previous invalid upstream interface are transferred to the new valid upstream interfaces which are chosen among the default upstream interfaces of the corresponding downstream nodes. The choice may be made based on the predefined policy (e.g., the interface priority, the number of listeners, the lowest IP address). An alternative may be applied by the MUIIMP device itself according to the traffic monitored or some strategies configured by the operator.

## 4. Security Considerations

To be done.

## 5. References

- [1] B. Fenner, H. He, B. Haberman and H. Sandick. "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.
- [2] Cain, B., Deering, S., Kouvelas, I., Fenner, B. and A. Thyagarajan, "Internet Group Management Protocol, Version3", RFC 3376, October 2002.
- [3] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.

- [4] Hong-Ke Zhang, Zhi-Wei Yan, Shuai Gao, et al., "Multicast Source Mobility Support in PMIPv6 Network", draft-zhang-multimob-msm-03, July 2011.
- [5] T C. Schmidt, S. Gao, H. Zhang, M. Waehlich, "Mobile Multicast Sender Support in Proxy Mobile IPv6 (PMIPv6) Domains", draft-ietf-multimob-pmipv6-source-02, October 2012.
- [6] T. Schmidt, M. Waehlich and S. Krishnan. "Base Deployment for Multicast Listener Support in Proxy Mobile IPv6 (PMIPv6) Domains", RFC 6224, April 2011.
- [7] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2000.

Authors' Addresses

Hong-Ke Zhang, Shuai-Gao, Bo-Hao Feng, Li-Li Wang  
National Engineering Lab for NGI Interconnection Devices  
Beijing Jiaotong University, China

Phone: +861051684274  
Email: [hkzhang@bjtu.edu.cn](mailto:hkzhang@bjtu.edu.cn)  
[shgao@bjtu.edu.cn](mailto:shgao@bjtu.edu.cn)  
[11111021@bjtu.edu.cn](mailto:11111021@bjtu.edu.cn)  
[liliwang@bjtu.edu.cn](mailto:liliwang@bjtu.edu.cn)

Thomas C. Schmidt  
HAW Hamburg  
Berliner Tor 7  
Hamburg 20099  
Germany

Email: [schmidt@informatik.haw-hamburg.de](mailto:schmidt@informatik.haw-hamburg.de)  
URI: <http://inet.cpt.haw-hamburg.de/members/schmidt>

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: August 26, 2013

W. Zhou  
cisco Systems  
February 22, 2013

VRRP PIM Interoperability  
draft-zhou-pim-vrrp-01.txt

Abstract

This document introduces VRRP Aware PIM, a redundancy mechanism for the Protocol Independent Multicast (PIM) to interoperate with Virtual Router Redundancy Protocol (VRRP). It allows PIM to track VRRP state and to preserve multicast traffic upon failover in a redundant network with virtual routing groups enabled.

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 26, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.



## Table of Contents

1. Introduction . . . . .	3
2. Tracking and Failover . . . . .	4
3. PIM Assert Metric Auto-Adjustment . . . . .	5
4. DF Election for BiDir Group . . . . .	6
5. Tracking Multiple VRRP Groups on an Interface . . . . .	7
6. Support of HSRP . . . . .	8
7. Security Considerations . . . . .	9
8. Acknowledgments . . . . .	10
9. Informative References . . . . .	11
Author's Address . . . . .	12

## 1. Introduction

Virtual Router Redundancy Protocol (VRRP) [RFC5798] is a redundancy protocol for establishing a fault-tolerant default gateway. The protocol establishes a framework between network devices in order to achieve default gateway failover if the primary gateway becomes inaccessible .

PIM has no inherent redundancy capabilities and its operation is completely independent of VRRP group states. As a result, IP multicast traffic is forwarded not necessarily by the same device as is elected by VRRP. The VRRP Aware PIM feature provides consistent IP multicast forwarding in a redundant network with virtual routing groups enabled.

In a multi-access segment (such as LAN), PIM designated router (DR) election is unaware of the redundancy configuration, and the elected DR and VRRP master router (MR) may not be the same router. In order to ensure that the PIM DR is always able to forward PIM Join/Prune message towards RP or FHR, the VRRP MR becomes the PIM DR (if there is only one VRRP group). PIM is responsible for adjusting DR priority based on the group state. When a failover occurs, multicast states are created on the new MR elected by the VRRP group and the MR assumes responsibility for the routing and forwarding of all the traffic addressed to the VRRP virtual IP address. This ensures the PIM DR runs on the same gateway as the VRRP MR and maintains mroute states. It enables multicast traffic to be forwarded through the VRRP MR, allowing PIM to leverage VRRP redundancy, avoid potential duplicate traffic, and enable failover, depending on the VRRP states in the device.

## 2. Tracking and Failover

With VRRP Aware PIM enabled, PIM listens to the state change notifications from VRRP and automatically adjusts the priority of the PIM DR based on the VRRP state, and ensures VRRP MR (if there is only one VRRP group) becomes the DR of the LAN. If there are multiple VRRP groups, the DR is determined by user-configured priority.

PIM triggers communication between upstream and downstream devices upon failover in order to create mroute states on the new MR. Depending on the requirements, there are various implementation options:

- o PIM sends additional PIM Hello message using the VRRP virtual IP addresses as the source address for each active VRRP group when a device becomes VRRP Active. The PIM Hello will carry a new GenID in order to trigger other routers to respond to the failover. When a downstream device receives this PIM Hello, it will add the virtual address to its PIM neighbor list. The new GenID carried in the PIM Hello will trigger downstream routers to resend PIM Join messages towards the virtual address. Upstream routers will process PIM Join/Prunes (J/P) based on VRRP group state.
- o An alternative solution is to have all passive routers maintain mroute states and record the GenID of current MR. When a passive router becomes MR upon switchover, it uses the existing mroute states and the recorded MR GenID in its Hello message. This solution avoids resending PIM J/P upon switchover and eliminates the requirement of additional PIM Hello with virtual IP address.

If the J/P destination matches the VRRP group virtual address and if the destination device is in VRRP active state, the new MR processes the PIM Join because it is now the acting PIM DR. This allows all PIM Join/Prunes to reach the VRRP group virtual address and minimizes changes and configurations at the downstream routers side.

### 3. PIM Assert Metric Auto-Adjustment

It is possible that, after VRRP active switched from A to B; A is still forwarding multicast traffic which will result in duplicate traffic and PIM Assert mechanism will kick in. PIM Assert with redundancy is enabled.

- o If only one VRRP group, passive routers will send a large penalty metric preference (PIM\_ASSERT\_INFINITY - 1) and make MR the Assert winner.
- o If there are multiples VRRP groups configured on an interface, Assert metric preference will be (PIM\_ASSERT\_INFINITY - 1) if and only if all VRRP groups are in passive.
- o If there is at least one VRRP group is in Active, then original Assert metric preference will be used. That is, winner will be selected between routers using their real Assert metric preference with at least one active VRRP Group, just like no VRRP is involved.

#### 4. DF Election for BiDir Group

Change to DF offer/winner metric is handled similarly to PIM Assert handling with VRRP.

- o If only one VRRP group, passive routers will send a large penalty metric preference in Offer (`PIM_BIDIR_INFINITY_PREF- 1`) and make MR the DF winner.
- o If there are multiples VRRP groups configured on an interface, Offer metric preference will be (`PIM_BIDIR_INFINITY_PREF- 1`) if and only if all VRRP groups are in passive.
- o If there is at least one VRRP group is in Active, then original Offer metric preference to RP will be used. That is, winner will be selected between routers using their real Offer metric with at least one active VRRP Group, just like no VRRP is involved.

## 5. Tracking Multiple VRRP Groups on an Interface

User can configure PIM to track more than one VRRP groups on an interface. This allows other applications to exploit the PIM/VRRP interoperability to achieve various goals (e.g., load balancing). Since each VRRP groups configured on an interface could be in different states at any moment, the DR priority is adjusted. PIM Assert metric and PIM Bidir DF metric if and only if all VRRP groups configured on an interface are in passive (non-Active) states to ensure that interfaces with all-passive VRRP groups will not win in DR, Assert and DF election. In other words, DR, Assert, DF winner will be elected among the interfaces with at least one Active VRRP group.

## 6. Support of HSRP

Although there are differences between VRRP and Hot Standby Router Protocol (HSRP) [RFC2281] including number of backup (standby) routers, virtual IP address and timer intervals, the proposed scheme can also enable HSRP aware PIM with similar switchover and tracking mechanism described in this draft.

## 7. Security Considerations

The proposed tracking mechanism has no negative impact on security.



## 8. Acknowledgments

I would like to give a special thank you and appreciation to Stig Venaas for his ideas and comments in this draft.

## 9. Informative References

- [RFC2281] Li, T., Cole, B., Morton, P., and D. Li, "Cisco Hot Standby Router Protocol (HSRP)", RFC 2281, March 1998.
- [RFC5798] Nadas, S., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC 5798, March 2010.

Author's Address

Wei Zhou  
cisco Systems  
Tasman Drive  
San Jose, CA 95134  
USA

Email: [weizho2@cisco.com](mailto:weizho2@cisco.com)



Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: June 13, 2013

L. Zheng  
Huawei Technologies  
Z. Zhang  
Juniper Networks  
R. Parekh  
Cisco Systems  
December 10, 2012

Survey Report on PIM-SM Implementations and Deployments  
draft-zzp-pim-rfc4601-update-survey-report-00.txt

Abstract

This document provides supporting documentation to advance the Protocol Independent Multicast - Sparse Mode (PIM-SM) routing protocol from IETF Proposed Standard to Internet Standard.

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 June 13, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as

described in the Simplified BSD License.

## Table of Contents

1. Motivation . . . . .	3
1.1. Overview of PIM-SM . . . . .	3
1.2. RFC2026 and RFC6410 Requirements . . . . .	3
2. Survey on Implementations and Deployments . . . . .	4
2.1. Methodology . . . . .	4
2.2. Operator Responses . . . . .	4
2.2.1. Description of PIM Sparse-Mode deployments . . . . .	4
2.2.2. PIM Sparse-Mode deployment with other multicast technologies . . . . .	4
2.2.3. PIM Sparse-Mode RPs and RP Discovery mechanisms . . . . .	4
2.3. Vendor Responses . . . . .	5
2.3.1. RFC4601 and RFC2362 implementations . . . . .	5
2.3.2. Lack of (*,*,RP) and PMBR implementations . . . . .	5
2.3.3. Implementations of other features of RFC4601 . . . . .	5
2.4. Key Findings . . . . .	6
3. Security Considerations . . . . .	7
4. IANA Considerations . . . . .	8
5. Acknowledgements . . . . .	9
6. Appendix A. Questionnaire . . . . .	10
6.1. Appendix A.1 PIM Survey for Operators . . . . .	10
6.2. Appendix A.2 PIM Survey for Implementors . . . . .	11
7. References . . . . .	14
7.1. Normative References . . . . .	14
7.2. Informative References . . . . .	14
Authors' Addresses . . . . .	15

## 1. Motivation

### 1.1. Overview of PIM-SM

PIM-SM was first published as [RFC2117] in 1997 and then again as [RFC2362] in 1998. The protocol was classified as Experimental in both of these documents. The PIM-SM protocol specification was then rewritten in whole and advanced to Proposed Standard as [RFC4601] in 2006. Considering its multiple independent implementations developed and sufficient successful operational experience gained, the IETF has decided to advance the PIM-SM routing protocol to Internet Standard.

### 1.2. RFC2026 and RFC6410 Requirements

[RFC2026] defines the stages in the standardization process, the requirements for moving a document between stages and the types of documents used during this process. Section 4.1.2 of [RFC2026] states that: "The requirement for at least two independent and interoperable implementations applies to all of the options and features of the specification. In cases in which one or more options or features have not been demonstrated in at least two interoperable implementations, the specification may advance to the Draft Standard level only if those options or features are removed."

[RFC6410] updates the Internet Engineering Task Force (IETF) Standards Process defined in [RFC2026]. Primarily, it reduces the Standards Process from three Standards Track maturity levels to two. The second maturity level is a merger of Draft Standard and Standard as specified in [RFC2026]. Section 2.2 of [RFC6410] states that: "(1) There are at least two independent interoperating implementations with widespread deployment and successful operational experience. (3) There are no unused features in the specification that greatly increase implementation complexity."

Optional features which do not meet the foresaid criteria has been identified by the PIM Working Group and will be removed. This document intends to provide supporting documentation to advance the Protocol Independent Multicast - Sparse Mode (PIM-SM) routing protocol from IETF Proposed Standard to Draft Standard.

## 2. Survey on Implementations and Deployments

### 2.1. Methodology

A questionnaire had been issued by the PIM WG co-chairs and announced widely to the vendors and operational community to obtain information on PIM-SM implementations and deployments. The Survey concluded on 22nd Oct 2012. The responses will be kept strictly confidential and only combined results will be published. The raw questionnaire will be shown in Appendix A, and a detailed summary of the responses will be included in the following section.

### 2.2. Operator Responses

Nine operators responded to the survey. They are SWITCH, National Research Council Canada, South Dakota School of Mines and Technology, Motorola Solutions and five other anonymous operators.

#### 2.2.1. Description of PIM Sparse-Mode deployments

In the last fourteen years, PIM-SM has been deployed for a wide variety of applications: Campus, Enterprise, Research and WAN networks, Broadband ISP and Digital TV. There are five deployments based on [RFC4601] implementation and two on [RFC2362] implementations. PIM-SM for IPv6 has been deployed by three operators. Out of the nine operators, six have deployed PIM-SM implementations from multiple vendors.

Operators reported minor inter-operability issues and these were addressed by the vendors. There was no major inter-operability concern reported by the operators.

#### 2.2.2. PIM Sparse-Mode deployment with other multicast technologies

Except for one deployment of PIM Sparse-Mode with Multicast OSPF (MOSPF), all other operators have deployed PIM-SM exclusively. No operators acknowledged deployments of either (\*,\*,RP) or Pim Multicast Border Route (PMBR) for inter-connection between PIM Sparse-Mode and other multicast domains.

#### 2.2.3. PIM Sparse-Mode RPs and RP Discovery mechanisms

The number of Sparse-Mode RPs deployed by operators range from a few (up to sixteen) to a massively scaled number (four hundred). Both static configuration and Bootstrap Router (BSR) have been deployed as RP discovery mechanisms.

Anycast-RP has been deployed for RP redundancy. Two operator have



deployed Anycast-RP using MSDP and three operators deploy both MSDP and PIM-SM Anycast-RP. The best common practice seems to be to use static-RP configuration with Anycast-RP for redundancy.

### 2.3. Vendor Responses

Eight vendors have reported PIM Sparse-Mode implementations. They are XORP, Huawei Technologies, Cisco Systems, Motorola Solutions, Juniper Networks and three other anonymous vendors.

#### 2.3.1. RFC4601 and RFC2362 implementations

Four vendors have reported implementations based on [RFC4601] and two have implemented PIM Sparse-Mode based on [RFC2362]. Two implementations are hybrid.

Minor inter-operability issues have been addressed by the vendors over the years and no concern was reported by any vendor.

#### 2.3.2. Lack of (\*,\*,RP) and PMBR implementations

Most vendors have not implemented (\*,\*,RP) state as specified in [RFC4601] either due to lack of deployment requirements or due to security concerns. Similarly, most vendors have also not implemented PMBR due to lack of deployment requirements or because it was considered to be too complex and non-scalable.

Only one vendor, XORP, reported (\*,\*,RP) and PMBR implementation and they were implemented just because these were part of the [RFC4601] specification.

#### 2.3.3. Implementations of other features of RFC4601

Most vendors have implemented all of the following from [RFC4601] specifications:

- SSM
- Join Suppression
- Explicit tracking
- Register mechanism
- SPT switchover at last-hop router
- Assert mechanism

- Hashing of group to RP mappings

Some vendors do not implement explicit tracking and SSM.

#### 2.4. Key Findings

1. PIM Sparse-Mode has been widely implemented and deployed for different applications. The PIM Sparse-Mode protocol is sufficiently well specified in RFC 4601 resulting in inter-operable implementation deployed by operators.

2. There are no deployments and only one known implementation of (\*,\*,RP) and PMBR as specified in RFC 4601. Hence, it is necessary to remove these features from the specification as required by [RFC2026] and [RFC6410]

### 3. Security Considerations

This document does not directly affect the security of the Internet.

#### 4. IANA Considerations

This document makes no request of the IANA.

## 5. Acknowledgements

The authors would like to thanks Tim Chown and Bill Atwood who had helped to collect and anonymize the responses as the neutral third-party. Special thanks are also given to Alexander Gall, William F Maton Sotomayor, Steve Bauer, Sonum Mathur, Pavlin Radoslavov, Shuxue Fan, Sameer Gulrajani and to the anonymous responders.

## 6. Appendix A. Questionnaire

This appendix reproduces a questionnaire that was made available for operators and vendors to express their experience and considerations.

### 6.1. Appendix A.1 PIM Survey for Operators

#### Introduction:

PIM-SM was first published as RFC2117 in 1997 and then again as RFC2362 in 1998. The protocol was classified as Experimental in both of these documents. The PIM-SM protocol specification was then rewritten in whole and advanced to Proposed Standard as RFC4601 in 2006. Considering the multiple independent implementations developed and the successful operational experience gained, the IETF has decided to advance the PIM-SM routing protocol to Draft Standard. This survey intends to provide supporting documentation to advance the Protocol Independent Multicast - Sparse Mode (PIM-SM) routing protocol from IETF Proposed Standard to Draft Standard. (Due to RFC6410, now the intention is to progress it to Internet Standard. Draft Standard is no longer used.)

This survey is issued on behalf of the IETF PIM Working Group.

The responses will be collected by a neutral third-party and kept strictly confidential if requested in the response; only the final combined results will be published. Tim Chown and Bill Atwood have agreed to anonymize the response to this Questionnaire. They have a long experience with multicast but have no direct financial interest in this matter, nor ties to any of the vendors involved. Tim is working at University of Southampton, UK, and he has been active in the IETF for many years, including the mboned working group, and he is a co-chair of the 6renum working group. Bill is at Concordia University, Montreal, Canada, and he has been an active participant in the IETF pim working group for over ten years, especially in the area of security.

Please send questionnaire responses addressed to them both. The addresses are [tjc@ecs.soton.ac.uk](mailto:tjc@ecs.soton.ac.uk) and [william.atwood@concordia.ca](mailto:william.atwood@concordia.ca). Please include the string "RFC4601 bis Questionnaire" in the subject field.

Before answering the questions, please complete the following background information.

Name of the Respondent:

Affiliation/Organization:

Contact Email:

Provide description of PIM deployment:

Do you wish to keep the information provided confidential:

Questions:

- 1 Have you deployed PIM-SM in your network?
- 2 How long have you had PIM-SM deployed in your network? Do you know if your deployment is based on the most recent RFC4601?
- 3 Have you deployed PIM-SM for IPv6 in your network?
- 4 Are you using equipment with different (multi-vendor) PIM-SM implementations for your deployment?
- 5 Have you encountered any inter-operability or backward-compatibility issues amongst differing implementations? If yes, what are your concerns about these issues?
- 6 Have you deployed both dense mode and sparse mode in your network? If yes, do you route between these modes using features such as \*,\*,RP or PMBR?
- 7 To what extent have you deployed PIM functionality, like BSR, SSM, and Explicit Tracking?
- 8 Which RP mapping mechanism do you use: Static, AutoRP, or BSR?
- 9 How many RPs have you deployed in your network?
- 10 If you use Anycast-RP, is it Anycast-RP using MSDP (RFC 3446) or Anycast-RP using PIM (RFC4610)?
- 11 Do you have any other comments on PIM-SM deployment in your network?

## 6.2. Appendix A.2 PIM Survey for Implementors

Introduction:

PIM-SM was first published as RFC2117 in 1997 and then again as RFC2362 in 1998. The protocol was classified as Experimental in both of these documents. The PIM-SM protocol specification was then rewritten in whole and advanced to Proposed Standard as RFC4601 in 2006. Considering the multiple independent implementations developed

and the successful operational experience gained, the IETF has decided to advance the PIM-SM routing protocol to Draft Standard. This survey intends to provide supporting documentation to advance the Protocol Independent Multicast - Sparse Mode (PIM-SM) routing protocol from IETF Proposed Standard to Draft Standard. (Due to RFC6410, now the intention is to progress it to Internet Standard. Draft Standard is no longer used.)

This survey is issued on behalf of the IETF PIM Working Group.

The responses will be collected by a neutral third-party and kept strictly confidential if requested in the response; only the final combined results will be published. Tim Chown and Bill Atwood have agreed to anonymize the response to this Questionnaire. They have a long experience with multicast but have no direct financial interest in this matter, nor ties to any of the vendors involved. Tim is working at University of Southampton, UK, and he has been active in the IETF for many years, including the mboned working group, and he is a co-chair of the 6renum working group. Bill is at Concordia University, Montreal, Canada, and he has been an active participant in the IETF pim working group for over ten years, especially in the area of security.

Please send questionnaire responses addressed to them both. The addresses are tjc@ecs.soton.ac.uk and william.atwood@concordia.ca. Please include the string "RFC 4601 bis Questionnaire" in the subject field.

Before answering the questions, please complete the following background information.

Name of the Respondent:

Affiliation/Organization:

Contact Email:

Provide description of PIM implementation:

Do you wish to keep the information provided confidential:

Questions:

1 Have you implemented PIM-SM?

2 Is the PIM-SM implementation based on RFC2362 or RFC4601?

3 Have you implemented (\*,\*, RP) state of RFC4601? What is the



rationale behind implementing or omitting (\*,\*,RP)?

4 Have you implemented the PMBR as specified in RFC4601 and RFC2715?  
What is the rationale behind implementing or omitting PMBR?

5 Have you implemented other features and functions of RFC4601:

- SSM
- Join Suppression
- Explicit tracking
- Register mechanism
- SPT switchover at last-hop router
- Assert mechanism
- Hashing of group to RP mappings

6 Does your PIM-SM implementation support IPv6?

7 Have you encountered any inter-operability issues with other PIM implementations in trials or in the field?

8 Do you have any other comments or concerns about PIM-SM as specified in RFC4601?

## 7. References

### 7.1. Normative References

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

### 7.2. Informative References

- [RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.
- [RFC2117] Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., Jacobson, V., Liu, C., Sharma, P., and L. Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2117, June 1997.
- [RFC2362] Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., and V. Jacobson, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June 1998.
- [RFC6410] Housley, R., Crocker, D., and E. Burger, "Reducing the Standards Track to Two Maturity Levels", BCP 9, RFC 6410, October 2011.

Authors' Addresses

Lianshu Zheng  
Huawei Technologies  
China

Email: [vero.zheng@huawei.com](mailto:vero.zheng@huawei.com)

Zhaohui Zhang  
Juniper Networks  
USA

Email: [zzhang@juniper.net](mailto:zzhang@juniper.net)

Rishabh Parekh  
Cisco Systems  
USA

Email: [riparekh@cisco.com](mailto:riparekh@cisco.com)

