

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)



MBONED  
Internet-Draft  
Intended status: Informational  
Expires: April 25, 2014

W. Atwood  
B. Li  
Concordia University/CSE  
S. Islam  
United International University/CSE  
October 22, 2013

Architecture for IP Multicast Receiver Access Control  
draft-atwood-mboned-mrac-arch-00

Abstract

This document specifies the architecture of IP multicast receiver access control (MRAC). The interacting components, the protocols and the operations in MRAC system are specified in this document.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 25, 2014.

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 . . . . .	2
2. MRAC Architecture Overview . . . . .	3
3. Multicast Architecture . . . . .	5
3.1. IGMP/MLD . . . . .	5
3.2. Multicast Routing Protocol (MRP) . . . . .	6
4. AAA Architecture . . . . .	6
4.1. Diameter / RADIUS . . . . .	6
4.2. EAP . . . . .	6
4.3. PANA . . . . .	7
5. IP Security (IPsec) Architecture . . . . .	8
6. MRAC System Operation . . . . .	8
6.1. Mapping the Multicast and AAA Architectures to the Network Boxes . . . . .	8
6.2. EAP Exchanges . . . . .	9
6.3. PANA Exchanges . . . . .	10
6.4. Diameter/RADIUS Exchanges . . . . .	10
6.5. IPsec Exchanges . . . . .	10
6.6. IGMP/MLD Exchanges . . . . .	11
7. Security Considerations . . . . .	11
8. IANA Considerations . . . . .	11
9. Acknowledgements . . . . .	11
10. References . . . . .	11
10.1. Normative References . . . . .	11
10.2. Informative References . . . . .	12
Authors' Addresses . . . . .	13

## 1. Introduction

Group communication at the application level usually implies IP multicast at the network level. The use of IP multicast can only be justified in certain environments if it is possible to authenticate the receiving users, and verify their authorization to receive the multicast data stream. However, the design of IP multicast [RFC1112] ensures that there can be no relationship between the sender and the receiving users, i.e., the sender is not aware of the identity of the receivers (or even if there are any receivers at all). This can make it very difficult for the sender to generate any revenue from the receivers of IP multicast services.

An alternative to access control by the sender is access control by the Network Service Provider, based on policies provided (directly or indirectly) by the sender. [I-D.atwood-mboned-mrac-req] lists the requirements on a set of mechanisms that allow a Network Service Provider to act on behalf of a sender (since the Network Service Provider has access to information from the receiving user that the sender does not have access to) to meet the access control and



revenue generation goals, while remaining as independent as possible from the specific business model in use.

This document assumes that the various business models could be simplified into two assumptions as follows:

The receiving user that receives the multicast data possesses a "ticket", which contains a description of the multicast group to be joined and whose validity can be demonstrated.

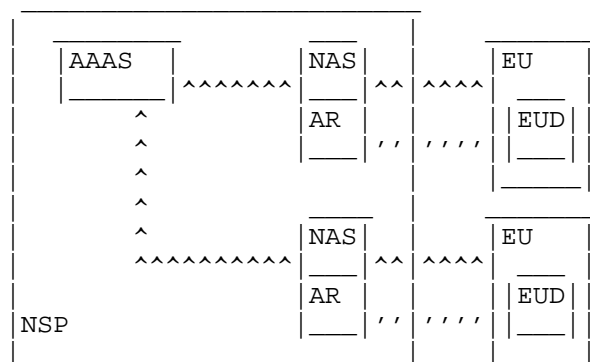
The Network Service Provider that delivers the multicast data possesses the "policies" that contain a description of how to validate the "ticket" of a receiving user.

[I-D.atwood-mboned-mrac-req] has presented how to fulfill the two assumptions in our business models.

This document proposes a Multicast Receiver Access Control (MRAC) architecture that satisfies all the requirements in [I-D.atwood-mboned-mrac-req]. In this draft, the above two assumptions are used so that the business model in use does not have to be considered.

## 2. MRAC Architecture Overview

The MRAC system has the interacting components as shown in Figure 1. A brief description of the components is as follows:



^^^^ Application-level access control flow  
 ' ' ' ' Network-level access control flow  
 EU End User  
 EUD End User Device  
 NSP Network Service Provider

AAAS	Authentication, Authorization and Accounting Server
AR	Access Router
NAS	Network Access Server

Figure 1: MRAC Architecture

End User (EU):

A subscriber who wishes to receive multicast data delivered in a multicast group. The EU possesses a "ticket" to verify his/her subscription for a specific group. However, the way how to distribute the ticket to the EU is dependent on the business model so that it is out of scope for this draft.

End User Device (EUD):

A device, connected to the Network Service Provider via one or more technologies, operated by an End User.

Network Service Provider (NSP):

An organization that delivers the multicast content to the End User Device and implements the access control of the receiving End Users for multicast groups. The NSP employs various devices to fulfill its services.

AAA Server (AAAS):

A device that manages authentication, authorization and accounting services for multicast groups in NSP. The AAAS possesses policies to validate the EU's tickets. However, the way how to distribute the policies to the AAAS is dependent on the business model so that it is out of scope for this draft.

Access Router (AR):

A router, close to the EU, which is responsible for adjudicating the access rights to the network and the multicast groups in the NSP.

Network Access Server (NAS):

The enforcement point for managing authentication, authorization and accounting services for multicast groups in the NSP. Normally the NAS is co-located with the Access Router.

The operations among the interacting components are generalized into two levels as follows:

Application-level operations:

The EU interacts with the NAS to request joining the group to which he/she has subscribed. The EU's ticket is carried in the request. The NAS consults the AAAS for the EU's request. The AAAS uses the policies to validate the ticket so as to authenticate the EU for the network and to authorize the EU for the multicast group. Then the AAAS returns the authentication and authorization result to the NAS. If the EU is authorized, some cryptographic materials derived from the ticket are also provided to the NAS by the AAAS. Moreover, the accounting information for the authorized EU is also exchanged between the NAS and the AAAS.

Network-level operations:

the EU interacts with the AR to join the data distribution tree that is delivering his/her subscription content. The exchanges between the EU and the AR at the network level are secured by the cryptographic materials derived from those used at the application level, to couple the access control for the two levels.

### 3. Multicast Architecture

#### 3.1. IGMP/MLD

Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) have been standardized by the IETF for IPv4 and IPv6 systems (host or router) to inform the neighbouring multicast router(s) about the multicast group memberships of these systems.

In IGMP, an IPv4 system sends a join or leave message (through Membership Report) when it wants to join or leave a multicast group (or some specific sources of a group). All multicast routers that are directly connected to the IPv4 system receive the Membership Report to learn which multicast groups are of interest to the IPv4 systems. Among these multicast routers, only one will be elected as "Querier". Its role is to query IPv4 systems about their interest in multicast groups by sending Membership Query. The other multicast routers are called Non-Queriers; they just receive the Membership Report messages from the IPv4 system and the Membership Query message from the Querier.

MLD is a similar protocol but it is used by IPv6 systems.

The details of the IGMP/MLD architecture and operation are specified in [RFC3376] and [RFC3810].

### 3.2. Multicast Routing Protocol (MRP)

The multicast routing protocol (typically PIM-SM) builds a multicast tree among routers to distribute multicast data to receivers.

One of the receiver's neighbouring routers is elected as the designated router (DR) or group designated router (GDR). On receiving the IGMP/MLD join message, DR/GDR will be grafted to the multicast data distribution tree on behalf of the neighbouring receiver. On receiving an IGMP leave Message, DR/GDR will be pruned from the data distribution tree if no other neighbouring receivers are interested in the group.

The details of PIM-SM architecture and operation are specified in [RFC4601].

## 4. AAA Architecture

### 4.1. Diameter / RADIUS

AAA protocols are used to support AAA communication between a AAA client and AAA server(s). RADIUS and Diameter are the two AAA protocols that the IETF has standardized.

Remote Authentication Dial In User Service (RADIUS) is a protocol for carrying information related to authentication, authorization, and accounting between a RADIUS Client and RADIUS Server. A RADIUS Client is responsible for passing user information to designated RADIUS Servers, and then acting on the response that is returned. RADIUS Servers are responsible for receiving user connection requests, authenticating the user, and then returning all configuration information necessary for the client to deliver service to the user.

Diameter is the successor of RADIUS. The Diameter base protocol provides a AAA framework. The Diameter applications, such as NASREQ and Mobile IPv4, specify how to use the base protocol within the context of their applications.

The details of Diameter / RADIUS architecture and operation are specified in [RFC6733] and [RFC2865]

### 4.2. EAP

Extensible Authentication Protocol (EAP) provides an authentication framework for support of multiple authentication methods.

An EAP exchange runs between an authenticator and a peer. The authenticator as an initiator uses one or more EAP methods in sequence to authenticate the peer.

Rather than requiring the authenticator to support authentication methods, EAP permits the use of a backend authentication server, which implements EAP methods, with the authenticator acting as a pass-through. In this case, a backend authentication server is connected with the authenticator. The actual authentication will be performed by the backend authentication server. The authenticator forwards EAP packets received from the peer to the backend authentication server; packets received from the backend authentication server are forwarded to the peer.

EAP does not run directly over the IP layer. The PANA protocol (Section 4.3) and the AAA protocol (Section 4.1) will be used to carry the EAP packets.

The details of the EAP architecture and operation are specified in [RFC3748].

#### 4.3. PANA

Protocol for carrying Authentication for Network Access (PANA) is a network access authentication protocol that works as an EAP lower layer for transmitting EAP packets. PANA carries EAP authentication methods (encapsulated inside EAP packets) between a PANA Client (PaC) and a PANA Authentication Agent (PAA) in the access network.

The PaC, as the client implement of PANA, interacts with the PAA, as the server implement of PANA, in the authentication process using the PANA protocol. PAA consults an Authentication Server (AS) for authentication and authorization of a PaC. If the AS resides on the same node as the PAA, an API is sufficient for this interaction. When the PAA is separated from the AS, a AAA protocol (e.g., Diameter) will be used for their communication. The AS is a conventional backend AAAS that terminates the EAP and the EAP methods.

A PANA Enforcement Point (EP) allows (blocks) data traffic of an authorized (unauthorized) PaC. When the PAA and EP reside on the same node, they use an API for communication; otherwise, a protocol (e.g., SNMP) is required.

The details of the PANA architecture and operation are specified in [RFC5191] and [RFC5193].

## 5. IP Security (IPsec) Architecture

The IP security (IPsec) architecture is designed to provide security services for traffic at the IP layer. Most of the security services are provided through use of two traffic security protocols, the Authentication Header (AH)[RFC4302] and the Encapsulating Security Payload (ESP)[RFC4303], and through the use of cryptographic key management procedures and protocols.

A security association (SA) is created in both sender and receiver. It is a simplex "connection" that affords security services to the traffic carried by it. The sender and receiver maintain their local Security Association Database (SAD) to record their SAs.

In the unicast case, the SAs could be dynamically negotiated between the sender and the receiver using IKEv2 [RFC5996]. In contrast, in the multicast case, a Group Controller / Key Server (GCKS) is responsible for distribution of the group SAs (GSA) to all the group members (GM) in the same group.

The details of the IPsec architecture and its extension for multicast are specified in [RFC4301] and [RFC5374].

## 6. MRAC System Operation

The MRAC architecture is composed from individual elements drawn from the pieces outlined in Sections 3, 4, and 5. In this way, it is possible to meet the requirements as listed in [I-D.atwood-mboned-mrac-req].

### 6.1. Mapping the Multicast and AAA Architectures to the Network Boxes

In order to meet the requirements in [I-D.atwood-mboned-mrac-req], all the functional entities that are applied in the multicast routing, AAA and IPsec architectures are mapped into the participants in our MRAC architecture as shown in Table 1.

The EU in MRAC system maps the IPv4/IPv6 system in IGMP/MLD, the receiver in multicast routing protocol, PaC in PANA and EAP peer in EAP.

The NAS in MRAC system maps the Diameter/RADIUS Client in Diameter/RADIUS, PAA in PANA and EAP Authenticator in EAP.

The AAAS in MRAC system maps the Diameter/RADIUS Server in Diameter/RADIUS, EAP backend authenticator server in EAP.

The AR in MRAC system maps the multicast router in IGMP/MLD including Querier and Non-Querier, the EP in PANA. Moreover, the AR who wins in DR/GDR election maps the DR/GDR in multicast routing protocol.

The EU and AR also map the GMs in IPsec. It depends on the specific packet whether the sender is EU or AR. A special AR, called Querier, may map the GCKS in IPsec.

Roles	EU	NAS	AAAS	AR
IPv4/IPv6 systems in IGMP	x			
multicast routers in IGMP				x
receiver in MRP	x			
DR/GDR in MRP				x
Diameter/RADIUS Client		x		
Diameter/RADIUS Server			x	
PaC in PANA	x			
PAA in PANA		x		
AS in PANA			x	
EP in PANA				x
EAP peer in EAP	x			
EAP authenticator in EAP		x		
backend authentication server in EAP			x	
GCKS in IPsec				x
GM in IPsec	x			x

Table 1: Mapping the Multicast and AAA Architectures to the Network Boxes

## 6.2. EAP Exchanges

The EAP runs between an EU and an NAS, where the NAS can act as a pass-through, and an AAAS is connected with the NAS. The policy is used to determine whether actual authentication is performed by the AAAS or the NAS.

In MRAC, it is recommended that the NAS use EAP-FAST as the EAP authentication method to authenticate the EU. The EU carries a token of his/her ticket in EAP-FAST method. The token includes the user information and group information to make the NAS or AAAS to do the authentication and authorization decision for the EU.

### 6.3. PANA Exchanges

PANA carries the EAP-FAST method (encapsulated inside EAP packets) between an EU and an NAS in the access network. An EU may be configured with an IP address of the NAS as its PAA or dynamically discover it using the default method of DHCP.

An EU, on receiving a request to join a multicast group while no GSAs have been established for the group, initiates a PANA exchange with an NAS as its PAA. During the PANA session, the NAS authenticates and authorizes the EU. In addition, the re-authentication and re-authorization may be required if necessary. The NAS would consult the AAAS to perform the actual authentication if it is a pass-through in the EAP exchange. The communication between NAS and AAAS is implemented by Diameter/RADIUS exchanges explained in the next subsection. Moreover, the NAS updates the filter state of the EU according to the authorization result and provides the authorization result and authorization attributes (mainly referring to a key derived from the EAP-FAST method) to the AR(s) connected directly to the EU.

### 6.4. Diameter/RADIUS Exchanges

When the NAS is a pass-through, the Diameter / RADIUS exchanges are used between NAS and AAAS to carry information related to authentication, authorization, and accounting.

In the Diameter / RADIUS exchanges, the NAS is responsible for passing the EAP-FAST method (carrying the token of the EU's ticket) to the AAAS, and then acting on the response that is returned. The AAAS is responsible for authenticating and authorizing the EU, and then returning the result and attributes (mainly including a key derived from the EAP-FAST method) for the EU to the NAS.

### 6.5. IPsec Exchanges

IPsec is used to enforce the receiver access control at the network level. A protocol is needed to create IPsec GSAs and then distribute them to authorized EUs and their ARs dynamically. The protocol may be very similar to the two existing IETF protocols, GDOI and G-IKEv2. In MRAC, the Querier will become the GCKS in the network segment. The authorized EUs and the other ARs interact with their Querier. The Querier is responsible for checking the authorized attributes of the EUs. Then the Querier creates and distributes IPsec GSAs to the EUs and their ARs in the same network segment.



However, the two existing IETF protocols (GDOI and G-IKEv2) do not satisfy all the requirements for IPsec GSA creation and distribution in MRAC. A new protocol has been drafted for use by MRAC.

#### 6.6. IGMP/MLD Exchanges

IGMP/MLD is running among the EUs and their ARs. According to [RFC3376] and [RFC3810], the EU uses the message of Membership Report to report the current multicast reception status to its ARs. The specific AR, called Querier, uses the message of Membership Query to query the multicast reception status of its EUs.

However, in order to enforce the receiver access control at the network level, the ARs and the EUs would filter out the received Membership Report and Membership Query messages using their local IPsec systems. The message of Membership Report that reports the reception status of the subscribed group would be protected by IPsec GSAs whose source address is the EU's address and whose destination address is the address of the reported subscribed group. Also, the Membership Query message that queries the reception status of the subscribed group would be protected by IPsec GSAs whose source address is the Querier's address and whose destination address is the address of the queried subscribed group.

In order to utilize the IPsec system, IGMP and MLD must be extended. However, the extension will be very small.

#### 7. Security Considerations

TBD

#### 8. IANA Considerations

This draft has no actions for IANA

#### 9. Acknowledgements

#### 10. References

##### 10.1. Normative References

[I-D.atwood-mboned-mrac-req]  
william.atwood@concordia.ca, w., Islam, S., and B. Li,  
"Requirements for IP Multicast Receiver Access Control",  
draft-atwood-mboned-mrac-req-00 (work in progress),  
October 2013.

- [RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, August 1989.
- [RFC2865] Rigney, C., Willens, S., Rubens, A., and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)", RFC 2865, June 2000.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [RFC3748] Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., and H. Levkowetz, "Extensible Authentication Protocol (EAP)", RFC 3748, June 2004.
- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC5191] Forsberg, D., Ohba, Y., Patil, B., Tschofenig, H., and A. Yegin, "Protocol for Carrying Authentication for Network Access (PANA)", RFC 5191, May 2008.
- [RFC5193] Jayaraman, P., Lopez, R., Ohba, Y., Parthasarathy, M., and A. Yegin, "Protocol for Carrying Authentication for Network Access (PANA) Framework", RFC 5193, May 2008.
- [RFC5374] Weis, B., Gross, G., and D. Ignjatic, "Multicast Extensions to the Security Architecture for the Internet Protocol", RFC 5374, November 2008.
- [RFC6733] Fajardo, V., Arkko, J., Loughney, J., and G. Zorn, "Diameter Base Protocol", RFC 6733, October 2012.

## 10.2. Informative References

- [RFC4302] Kent, S., "IP Authentication Header", RFC 4302, December 2005.
- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, December 2005.

[RFC5996] Kaufman, C., Hoffman, P., Nir, Y., and P. Eronen,  
"Internet Key Exchange Protocol Version 2 (IKEv2)", RFC  
5996, September 2010.

#### Authors' Addresses

William Atwood  
Concordia University/CSE  
1455 de Maisonneuve Blvd, West  
Montreal, QC H3G 1M8  
Canada

Phone: +1(514)848-2424 ext3046  
Email: [william.atwood@concordia.ca](mailto:william.atwood@concordia.ca)  
URI: <http://users.encs.concordia.ca/~bill>

Bing Li  
Concordia University/CSE  
1455 de Maisonneuve Blvd, West  
Montreal, QC H3G 1M8  
Canada

Email: [leebingice@gmail.com](mailto:leebingice@gmail.com)

Salekul Islam  
United International University/CSE  
House 80, Road 8/A, Mirza Golam Hafiz Road  
Dhanmondi, Dhaka 1209  
Bangladesh

Email: [salekul@cse.uiu.ac.bd](mailto:salekul@cse.uiu.ac.bd)

MBONED Working Group  
Internet-Draft  
Intended status: Informational  
Expires: April 24, 2014

W. Atwood  
Concordia University/CSE  
S. Islam  
United International University  
B. Li  
Concordia University/CSE  
October 21, 2013

Requirements for IP Multicast Receiver Access Control  
draft-atwood-mboned-mrac-req-00

Abstract

IP multicast offers no facilities for receiver access control or accounting. This document explores the requirements for such facilities.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 24, 2014.

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.

## Table of Contents

1. Introduction . . . . .	2
2. Previous Work . . . . .	4
3. Reference Architecture . . . . .	6
4. Requirements on the Solution . . . . .	10
4.1. Application-level constraints . . . . .	10
4.1.1. Authenticating and Authorizing Multicast End Users . . . . .	10
4.1.2. Group Membership and Access Control . . . . .	10
4.1.3. Independence of Authentication and Authorization Procedures . . . . .	11
4.1.4. Re-authentication and Re-authorization . . . . .	11
4.1.5. Accounting . . . . .	11
4.1.6. Multiple Sessions on One Device . . . . .	11
4.1.7. Multiple Independent Sessions on a LAN . . . . .	11
4.1.8. Application level interaction must be secured . . . . .	11
4.2. Network Level Constraints . . . . .	12
4.2.1. Maximum Compatibility with MLD and IGMP . . . . .	12
4.2.2. Minimal Modification to MLD/IGMP . . . . .	12
4.2.3. Multiple Network Level Joins for End User Device . . . . .	12
4.2.4. NSP Representative Differentiates Multiple Joins . . . . .	12
4.2.5. Network-level Interaction must be secured . . . . .	12
4.3. Interaction Constraints . . . . .	12
4.3.1. Coupling of Network and Application Level Controls . . . . .	12
4.3.2. Separation of Network Access Controls from Group Access Controls . . . . .	13
5. Security Considerations . . . . .	13
6. IANA Considerations . . . . .	13
7. Acknowledgements . . . . .	13
8. References . . . . .	13
8.1. Normative References . . . . .	13
8.2. Informative References . . . . .	14
Authors' Addresses . . . . .	16

## 1. Introduction

When using group communication at the application level, there is a variety of ways that the subscribers to a group (the End Users) can be managed. Encryption can be used at this level to secure the group data, i.e., to prevent a non-subscribing End User from interpreting the resulting group data as they are delivered.

When an End User joins an application-level group, this normally implies that the End User Device will join the corresponding network-level IP multicast group. The procedure for effecting this join, as defined in [RFC1112], is an open one:

- o a request is made by the receiving host, using MLD (IPv6) [RFC3810] or IGMP (IPv4) [RFC3376],
- o the Access Router that receives the request is required to use the multicast routing protocol (typically PIM-SM [RFC4601]) to graft itself to the network-level multicast data distribution tree.

This "unconditional join" implies that there is no access control at the network level, i.e., it is not possible to prevent an arbitrary End User Device from asking that the multicast data stream be delivered to it.

The unconditional construction of the data distribution tree is thus entirely receiver-driven, with the result that there is no relationship between the sender and the receiver(s), i.e., the sender is not aware of the identity of the receivers (or even if there are any receivers at all).

This can make it very difficult for the Content Provider to generate any revenue from the receivers of IP multicast services.

There are some environments where sufficient access control to the multicast data stream can be achieved because of the physical characteristics of the delivery medium (e.g., DSL links, point-to-point links).

There are some environments where access control is undesired or irrelevant (e.g., internal corporate distribution, subscriber controlled by a set-top box).

There are some environments where the use of multicast data distribution could result in resource savings (for the Content Provider and/or the Network Service Provider), but the Network Service Provider is reluctant to use this technology because of the inability to correlate the receiving End Users with the service being delivered, which makes it very difficult for the Network Service Provider to derive any revenue from the multicast stream.

Access control can be viewed at two levels: the application level and the network level. At the application level, an End User will obtain permission to subscribe to a group session. This permission will contain at least two components: a description of how the session is to be accessed and a certification that the End User is authorized to access the session.

The certification will be presented at the application level, and if it is valid the End User will be permitted to join the group.

At the network level, the session descriptor will be used to issue the network level join, which allows the session data to flow to the end user host.

To prevent the end user from presenting an arbitrary session descriptor, it is necessary to coordinate the application level join and the network level join. Two possible ways of achieving the necessary coordination are:

[Solution 1] Carry the application level rights certification in an extended network level join exchange;

[Solution 2] Provide separate application level join and network level join functions, along with a method for explicitly coordinating them.

Effective access control must be secured. It is not meaningful to implement access control without also ensuring that the party making the request for access (i.e., the End User) is authenticated. Since the network-level request is made using MLD/IGMP, this implies that the MLD/IGMP exchanges must also be secured.

The overall goal of this work is to list the requirements on a set of mechanisms that allow the Network Service Provider to act on behalf of the Content Provider (since the Network Service Provider has access to information from the End User that the Content Provider does not have access to) to meet the access control and revenue generation goals, while remaining as independent as possible from the specific business model in use.

## 2. Previous Work

Several pieces of the solution have received significant attention in recent years.

The problem of security and key management for application-level groups has been explored by the Multicast Security (MSEC) working group, and a framework devised [RFC3740].

The use of AAA protocols (RADIUS [RFC2865], Diameter [RFC3588]) to manage network-level access has been standardized. The approach outlined in this document is based on the observation that the AAA protocols (especially Diameter) can be extended to permit controlling access to application-level groups.

Some requirements for "well-managed" multicast have been stated in [I-D.ietf-mboned-maccnt-req], and a framework for satisfying these requirements with the help of AAA functionality has been described in

[I-D.ietf-mboned-multiaaaa-framework]. These documents suggest various business models for the interaction with the End User(s), with (potentially) separated functions corresponding to the Content Provider and the Network Service Provider.

The requirements document [I-D.ietf-mboned-maccnt-req] gives general requirements for authentication, authorization, accounting and Quality of Service (QoS) control. It assumes that the required goals can be achieved by integrating AAA with a multicast Content Distribution System, with MLD/IGMP at the edge of the network. The framework document [I-D.ietf-mboned-multiaaaa-framework] presents a basic AAA enabled model as well as an extended fully enabled model with resource and admission control coordination.

The approach of extending IGMP to carry authentication information has been proposed for a number of years [I-D.irtf-gsec-igmpv3-security-issues], [I-D.irtf-gsec-smrac], [I-D.he-magma-igmpv3-auth], [I-D.coan-hasm], [I-D.hayashi-igap].

Van Moffaert [I-D.irtf-gsec-igmpv3-security-issues] has proposed a mechanism for securing the IGMPv3 packets using IPsec. The IGMP [RFC3376] specification suggests the use of IPsec with Authentication Header (AH) [RFC4302] to secure the packet exchanges, and notes certain limitations on its use. The MLD [RFC3810] specification is silent on the issue of securing the packets.

A receiver access control architecture has been proposed in [MulticastReceiver] and [MulticastPANA]. In addition to the Network Service Provider and Content Provider of the "well-managed multicast" model, it incorporates the concepts of a Merchant (to offer the available services to the End User) and a Financial Institution (to verify the ability of the End User to pay for the desired services).

A sender access control architecture has been proposed in [MulticastSender].

[I-D.liu-mboned-mldauth-ps] provides additional requirements for the case where the End User device is mobile. [MulticastMobile] provides a solution for the issue of device mobility, using the EAP Reauthorization Protocol [RFC6696].

As an extensive mechanism for QoS management already exists [RSVP], this part of the problem will be considered to be out-of-scope for this document.

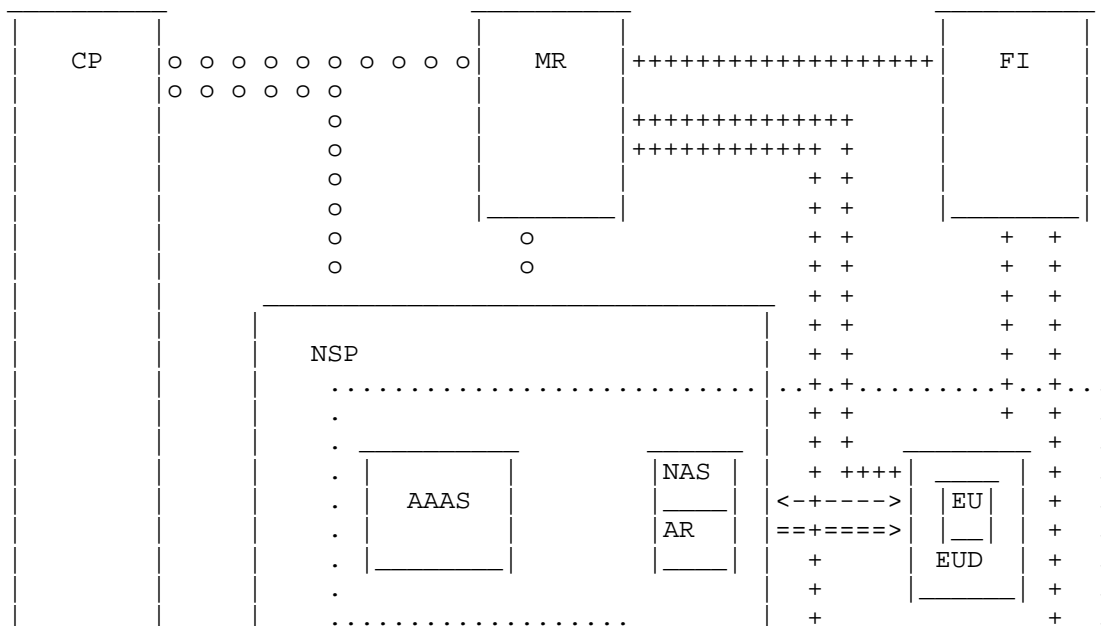
Finally, work is under way on securing the network routing infrastructure [RFC6862] [RFC6518]. In particular, securing of the exchanges between adjacent PIM-SM routers is specified in [RFC5796].

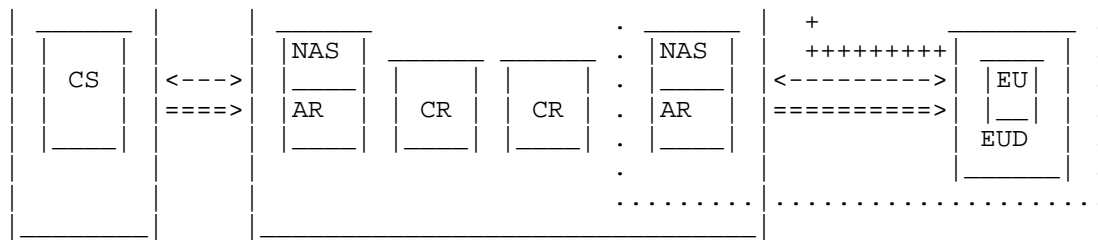


However, one key piece is missing. It is necessary to authenticate and authorize receiving users and to correlate their right to access a group with the action of putting the data on that part of the network that is directly connected to the receiving host. These two actions must be done securely, to ensure the correctness of the authentication and authorization actions. As noted in the Introduction, there are two approaches to achieving the correlation: carry the application-level information in the network-level join message, or separate the two messages and ensure that they are correlated cryptographically. These two approaches will be explored in Section 4, and the choice of one of them will be justified. Ensuring that the network-level join is not done unless the application-level join is authorized also has the desirable side effect of minimizing the resource wastage that would result from delivering multicast traffic to devices whose End Users have no entitlement to receive them.

### 3. Reference Architecture

A system for the delivery of multicast data will have interacting components, which are illustrated in Figure 1 to facilitate discussion. Note that only the components that are inside the dotted line are in scope for this document. The components outside the dotted line are presented only to show how the inside components relate to the outside components.





o o o Policy flow  
 +++++ Purchase flow  
 <----> Access Control flow  
 ====> Data flow  
 ..... Scope of interest

CP Content Provider  
 CS Content Server  
 MR Merchant  
 FI Financial Institution  
 EU End User  
 EUD End User Device  
 NSP Network Service Provider  
 AAAS AAA Server  
 AR Access Router  
 NAS Network Access Server  
 CR Core Router

Figure 1: Reference architecture

A brief description of the components follows:

Content Provider (CP): A person or organization that creates content for distribution.

Content Server (CS): A device that distributes the content via multicast data distribution.

Merchant (MR): An organization that offers content from one or more Content Providers to End Users, to be delivered via the facilities of one or more Network Service Providers.

Financial Institution (FI): An organization that certifies that a particular End User is able to pay for content that has been ordered through a Merchant.

Network Service Provider (NSP): An organization that delivers content from a Content Server to End User Devices.

AAA Server (AAAS): A device for managing Authentication, Authorization and Accounting within the Network Service Provider.

Access router (AR): A routing device within the Network Service Provider, close to the End User Device, which is responsible for adjudicating access rights to the network.

Network Access Server (NAS) The enforcement function for managing Authentication, Authorization and Accounting within the Network Service Provider. Normally co-located with the Access Router.

Core Router (CR): A routing device within the Network Service Provider that does not have any End User Device connected to it.

End User (EU): A subscriber who wishes to receive multicast data.

End User Device (EUD): A device, connected to the Network Service Provider via one or more technologies, operated by an End User.

These components illustrate separate functionalities. The functionalities may in fact be under separate administrative control, or they may be combined in various ways.

Since the end point of the NSP side of several interactions cannot be precisely determined until the detailed design is done, the term "NSP Representative" will be used in this document. A typical NSP Representative will be located on a router or other device that is "close" to the End User Device.

There are four kinds of information flow in Figure 1.

Policy flow: Exchange of policy information.

Purchase flow: The transactions related to subscribing to and paying for a group session.

Access Control flow: The presentation of authentication and authorization information.

Data flow: The delivery of the subscribed data stream.

The operation of the components and the exchange of information may be illustrated through the following example:

The Content Provider arranges to provide a live video multicast session for a football match. It contracts with the Merchant to act as its "sales agent", and provides relevant policies concerning the distribution of this particular content stream. The Merchant will

offer this content stream to interested subscribers (the End Users), using any available mechanism (in this case, its website, [www.mcast-football.com](http://www.mcast-football.com)). When an End User (Alice) subscribes to the content, the Merchant will verify with the Financial Institution that Alice is able to pay. Depending on the nature of the relationship among the Merchant, Alice, and her Financial Institution, the payment may be taken immediately, or it may be deferred to some point after delivery of the subscribed stream. The Merchant then issues a "ticket" to Alice, containing the information to identify Alice and information to identify the content stream to which she has subscribed. This could have, for example, the following form:

- o A pair of (public and private) keys generated by the Merchant exclusively for Alice, plus the digital certificate that authenticates the identity of Alice and carries the public key of Alice, which is signed by the Merchant or any other well-known Certificate Authority
- o The multicast address (e.g., w.x.y.z:port) to which the data would be sent.
- o If required, a symmetric key to decrypt the multicast data. (Note that the encryption is optional (but likely). It is also unrelated to the Access Control features, and so out of scope for this document.)

Alice has a video client that will process the multicast address and request receipt of the video stream at the network level. However, Alice's right to receive the video stream must also be established (transparently to Alice) before she starts to receive the subscribed stream. The requirements on this verification form the core of the purpose of this document. If the verification is successful, the Access Router will be grafted onto the multicast data distribution tree within the Network Service Provider. The multicast content is streamed to the Network Service Provider at the appropriate time by the Content Server. The Network Service Provider will begin to stream the content to the End User Device once it becomes available.

Policies concerning the access to the data stream are exchanged between the Content Provider and the Merchant; they may also be exchanged between the Content Provider and the Network Service Provider. Policies concerning the validation of a "ticket" are exchanged between the Merchant and the Network Service Provider; these may depend in part on the policies that were received by the Merchant from the Content Provider. In total, the policies received by the Network Service Provider are expected to contain sufficient information that the AAAS will be able to validate a ticket without having to refer directly to the Content Provider.

#### 4. Requirements on the Solution

As noted in the Introduction, access control can be viewed at two levels: the application level and the network level. To ensure that permissions given at the application level are reflected in the corresponding network-level actions, it is necessary to coordinate them. Section 2 has outlined several proposals that have been made in the past for extensions to IGMP to achieve this coordination. However, these proposals each appear to be heavily tied to a particular version of IGMP, and so will be incompatible with future versions of MLD or IGMP. In addition, such proposals are in effect a new version of MLD/IGMP, and even after several years, IGMPv3 is not universally available in mainstream Operating Systems. This makes it more desirable to find a solution to the access control problem that does not require the presence of application-level access control information in MLD (or IGMP) packets. Thus, the approach labeled "Solution 1" in Section 1 is assumed in the following.

To allow for independent development of application-level mechanisms and network-level mechanisms, the requirements in this document are based on the assumption that a single method can be used for securing the MLD/IGMP exchanges, where the associated cryptographic parameters for this method are correlated with the authentication and authorization that has already been done at the application level.

This leads to a natural separation of the requirements into three categories: constraints on the application-level interactions, constraints on the network-level interactions, and constraints on the coordination between them.

##### 4.1. Application-level constraints

###### 4.1.1. Authenticating and Authorizing Multicast End Users

The design of IP multicast [RFC1112] ensures that there can be no relationship between the End Users and the Content Provider(s). The primary goal is therefore to establish an equivalent relationship between each End User and the associated NSP Representative.

###### 4.1.2. Group Membership and Access Control

Although specifications exist for encrypting the user data, thus ensuring that only legitimate users can decrypt these data, these specifications provide no way to ensure that the data distribution tree is not extended when a non-authorized receiving user makes a request to join the tree. Thus, "group membership" and "multicast receiver access control" have to be considered (and solved) as separate problems.

#### 4.1.3. Independence of Authentication and Authorization Procedures

There is a wide range of authentication and authorization procedures that may be desired by an Internet Service Provider, including some that may not yet be standardized. This implies the adoption of a very general framework for such procedures. If a general framework is used, then it is likely to be independent of the specific business model in use by the CP or the NSP.

#### 4.1.4. Re-authentication and Re-authorization

Several scenarios can cause a need for re-authentication and re-authorization:

- o When a user changes the group that he/she wishes to attach to;
- o When a user changes the access router used for connection (e.g., wireless roaming);
- o When a user changes the medium used for physical connectivity (e.g., cellular to wireless, etc.).

This implies the need for a general solution to the access control problem that facilitates re-authentication and re-authorization.

#### 4.1.5. Accounting

The fact of delivery of group data needs to be recorded, to enable revenue to be earned. This is only one of a range of accounting issues that may need to be addressed, which points to the need for a general solution that allows a range of accounting actions to be supported.

#### 4.1.6. Multiple Sessions on One Device

Since an End User may wish to join multiple groups simultaneously, it must be possible to associate multiple sessions with a single End User Device.

#### 4.1.7. Multiple Independent Sessions on a LAN

Since multiple devices on a LAN may have End Users who wish to join the session, it must be possible to differentiate these End Users on the LAN.

#### 4.1.8. Application level interaction must be secured

Mutual authentication of the NSP Representative and the End User must be possible.

#### 4.2. Network Level Constraints

##### 4.2.1. Maximum Compatibility with MLD and IGMP

The proposed solution should be compatible with all current versions of MLD and IGMP. It is important that a solution not be tied to the semantics or packet format of a particular version of MLD or IGMP.

##### 4.2.2. Minimal Modification to MLD/IGMP

The solution developed should minimize any alteration to the semantics and the packet layout of MLD and IGMP.

##### 4.2.3. Multiple Network Level Joins for End User Device

It has to be possible for an End User Device to issue multiple distinct network-level join requests. (This is implied by the constraint in the Application level.)

##### 4.2.4. NSP Representative Differentiates Multiple Joins

It has to be possible for the NSP Representative to manage multiple Network Level joins for a single shared medium. (This is implied by the constraint in the Application level.)

##### 4.2.5. Network-level Interaction must be secured

Mutual authentication of the NSP Representative and the End User must be possible.

#### 4.3. Interaction Constraints

##### 4.3.1. Coupling of Network and Application Level Controls

It is conceivable that a solution could be found for the above issues that would be based on standard network protocols and separate (proprietary or standard) group management protocols. For example, the key management and distribution protocol associated with the application-level group could have authentication as one of its features. However, the separation of the network-level controls from the application-level controls enables a significant class of security attacks. It is therefore important that control of access to the network resources and control of access to the application-level resources be strongly coupled. This implies that the method used to cryptographically secure the MLD/IGMP interactions should be

strongly coupled to the method used to ensure authentication and authorization at the application level. However, it does not imply that the application-level interaction should be responsible for securing the network-level access, or that the network-level access should carry application-level information.

#### 4.3.2. Separation of Network Access Controls from Group Access Controls

Access to the network is different from access to a group. As an example, the authorization to watch a particular video presentation may be associated with a specific family member, while the authorization to use the network connection may be associated with an entire family (or to anyone present in the house).

While existing AAA procedures are designed to control network level access, they would have to be extended (or alternatives found) if group access needs to be controlled.

#### 5. Security Considerations

TBD.

#### 6. IANA Considerations

This document has no actions for IANA.

#### 7. Acknowledgements

#### 8. References

##### 8.1. Normative References

- [RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, August 1989.
- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2865] Rigney, C., Willens, S., Rubens, A., and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)", RFC 2865, June 2000.



- [RFC3588] Calhoun, P., Loughney, J., Guttman, E., Zorn, G., and J. Arkko, "Diameter Base Protocol", RFC 3588, September 2003.
- [RFC3748] Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., and H. Levkowetz, "Extensible Authentication Protocol (EAP)", RFC 3748, June 2004.
- [RFC4302] Kent, S., "IP Authentication Header", RFC 4302, December 2005.
- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, December 2005.

## 8.2. Informative 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.
- [RFC3740] Hardjono, T. and B. Weis, "The Multicast Group Security Architecture", RFC 3740, March 2004.
- [RFC5296] Narayanan, V. and L. Dondeti, "EAP Extensions for EAP Re-authentication Protocol (ERP)", RFC 5296, August 2008.
- [I-D.ietf-mboned-maccnt-req]  
Hayashi, T., Satou, H., Ohta, H., He, H., and S. Vaidya, "Requirements for Multicast AAA coordinated between Content Provider(s) and Network Service Provider(s)", draft-ietf-mboned-maccnt-req-10 (work in progress), August 2010.
- [I-D.ietf-mboned-multiaaaa-framework]  
Satou, H., Ohta, H., Hayashi, T., Jacquenet, C., and H. He, "AAA and Admission Control Framework for Multicasting", draft-ietf-mboned-multiaaaa-framework-12 (work in progress), August 2010.
- [I-D.liu-mboned-mldauth-ps]  
Liu, Y., Sarikaya, B., and P. Yang, "MLDv2 User Authentication Problem Statement", draft-liu-mboned-mldauth-ps-00 (work in progress), February 2008.
- [I-D.irtf-gsec-igmpv3-security-issues]  
Paridaens, O. and A. Moffaert, "Security issues in Internet Group Management Protocol version 3 (IGMPv3)", draft-irtf-gsec-igmpv3-security-issues-01 (work in progress), March 2002.

- [I-D.draft-ishikawa-igmp-auth]  
Ishikawa, N., Yamanouchi, N., and O. Takahashi, "IGMP Extension for Authentication of IP Multicast Senders and Receivers ", draft-ishikawa-igmp-auth-01 (work in progress), August 1998.
- [I-D.irtf-gsec-smrac]  
He, H., "Simple Multicast Receiver Access Control", draft-irtf-gsec-smrac-00 (work in progress), November 2001.
- [I-D.he-magma-igmpv3-auth]  
He, H., "Upload Authentication Information Using IGMPv3", draft-he-magma-igmpv3-auth-00 (work in progress), November 2001.
- [I-D.coan-hasm]  
Coan, B., "HASM: Hierarchical Application-Level Secure Multicast", draft-coan-hasm-00 (work in progress), December 2001.
- [I-D.hayashi-igap]  
Hayashi, T., "Internet Group membership Authentication Protocol (IGAP)", draft-hayashi-igap-03 (work in progress), August 2003.
- [RFC6862] Lebovitz, G., Bhatia, M., and B. Weis, "Keying and Authentication for Routing Protocols (KARP) Overview, Threats, and Requirements", RFC 6862, March 2013.
- [RFC6518] Lebovitz, G. and M. Bhatia, "Keying and Authentication for Routing Protocols (KARP) Design Guidelines", RFC 6518, February 2012.
- [RFC5796] Atwood, W., Islam, S., and M. Siami, "Authentication and Confidentiality in Protocol Independent Multicast Sparse Mode (PIM-SM) Link-Local Messages", RFC 5796, March 2010.
- [RFC6696] Cao, Z., He, B., Shi, Y., Wu, Q., and G. Zorn, "EAP Extensions for the EAP Re-authentication Protocol (ERP)", RFC 6696, July 2012.
- [MulticastReceiver]  
Islam, S. and W. Atwood, "Multicast Receiver Access Control by IGMP-AC, Computer Networks, doi://10.1016/j.comnet.2008.12.005", January 2009.
- [MulticastSender]

Islam, S. and W. Atwood, "Sender Access and Data Distribution Control for Inter-domain Multicast Groups, Computer Networks, doi://10.1016/j.comnet.2010.01.006", October 2010.

[MulticastPANA]

Islam, S. and W. Atwood, "Multicast Receiver Access Control using PANA, 1st Taibah University International Conference on Computing and Information Technology (ICCIT 2012), Al-Madinah Al-Munawwarah, Saudi Arabia, pp. 816--821. ", March 2012.

[MulticastMobile]

Islam, S. and W. Atwood, "Receiver Access Control and Secured Handoff in Mobile Multicast using IGMP-AC, LCN 2008, pp. 411--418", November 2008.

Authors' Addresses

J. William Atwood  
Concordia University/CSE  
1455 de Maisonneuve Blvd, West  
Montreal, QC H3G 1M8  
Canada  
  
Phone: +1(514)848-2424 ext3046  
Email: [william.atwood@concordia.ca](mailto:william.atwood@concordia.ca)  
URI: <http://users.encs.concordia.ca/~bill>

Salekul Islam  
United International University  
House # 80, Road # 8/A  
Mirza Golam Hafiz Road  
Dhanmondi, Dhaka 1209  
Bangladesh

Email: [salekul@cse.uiu.ac.bd](mailto:salekul@cse.uiu.ac.bd)

Bing Li  
Concordia University/CSE  
1455 de Maisonneuve Blvd, West  
Montreal, QC H3G 1M8  
Canada

Email: [leebingice@gmail.com](mailto:leebingice@gmail.com)

PIM  
Internet-Draft  
Intended status: Standards Track  
Expires: May 08, 2014

W. Atwood  
B. Li  
Concordia University/CSE  
November 04, 2013

Secure Internet Group Management Protocol  
draft-atwood-pim-sigmp-00

Abstract

This document specifies a Secure Internet Group Management Protocol (SIGMP), which is an extension to IGMP to enforce receiver access control for secured multicast groups. In SIGMP, only the hosts operated by authorized end users are permitted to report their interest in secured groups. IPsec is used to filter the messages that report or query the interest in secured groups. SIGMP provides two working modes that are fully compatible with IGMP v2 and IGMP v3 respectively.

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 May 08, 2014.

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 . . . . .	2
1.1. Terminology . . . . .	3
1.2. Assumptions . . . . .	3
2. Overview of SIGMP . . . . .	4
3. Packet Format . . . . .	5
4. Router Operations . . . . .	5
4.1. Router Operations Compatible with IGMP v2 . . . . .	5
4.1.1. Router Operations for a Received Report . . . . .	6
4.2. Router Operations Compatible with IGMP v3 . . . . .	7
4.2.1. Router Operations on a Received Report . . . . .	7
5. Host Operations . . . . .	8
5.1. Host Operations Compatible with IGMP v2 . . . . .	9
5.1.1. Conditions for Unsolicited Report . . . . .	9
5.1.2. Host Operations for a Received Query . . . . .	9
5.2. Host Operations Compatible with IGMP v3 . . . . .	9
5.2.1. Host Operations for a Received General Query . . . . .	9
6. IANA Considerations . . . . .	9
7. References . . . . .	9
7.1. Normative References . . . . .	10
7.2. Informative References . . . . .	10
Authors' Addresses . . . . .	10

## 1. Introduction

The Internet Group Management Protocol (IGMP) is used by IPv4 systems (hosts and routers) to report their IP multicast group memberships to any neighboring multicast routers. There are two popular versions: IGMP v2, as specified in [RFC2236] and IGMP v3, as specified in [RFC3376]. However, both versions establish a fully "open" multicast network, where any host can join any multicast group as a recipient without receiver access control.

This document specifies a Secure Internet Group Management Protocol (SIGMP) working in a "hybrid" multicast network. In a hybrid network, multicast groups are classified into two categories: open groups and secured groups. Open groups refer to multicast groups that any host can join unconditionally as a receiver. Secured groups refer to multicast groups with receiver access control, e.g., only hosts operated by authenticated and authorized end users are permitted to join as receivers. SIGMP retains most mechanisms of IGMP and enforces receiver access control to secured groups in a multicast network. On the one hand, any host could report its

interest in open groups freely as in IGMP. On the other hand, only hosts operated by the authenticated and authorized end users are permitted to report their interest in secured groups.

Instead of a new specific mechanism, SIGMP uses IPsec [RFC4301] to implement receiver access control to secured groups at the IP layer. Some Security Associations (SAs) are created to secure the SIGMP packets that are used to report or query secured groups. The packets coming from the unauthorized hosts will be discarded by the IPsec subsystem if they are used to report or query interest in secured groups.

### 1.1. 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 RFC 2119 [RFC2119].

It is assumed that the reader is familiar with the defining documents for IGMP [RFC2236] and [RFC3376]. Unless otherwise noted, terms defined in these documents are used with the same meaning in this one.

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

open group: A multicast group without receiver access control. Any host can unconditionally join any open group as a receiver, e.g. the data in a open group can be received by any host.

secured group: A multicast group with receiver access control. Only hosts operated by authenticated and authorized end users are permitted to join a secured group as a receiver, e.g. the data in a secured group can only be received by hosts operated by authenticated and authorized end users.

### 1.2. Assumptions

In order to focus on the actions of group membership (e.g., joining and leaving groups), the following topics are assumed to be discussed elsewhere:

1. how to distinguish between secured groups and open groups;
2. how to authenticate and authorize the operators of the devices (hosts and routers);

3. how to distribute the necessary Security Associations to participant devices (hosts and routers).

The existence of the group property (secured or open) defines the hybrid nature of the environment in which SIGMP works. A variety of existing protocols (e.g., LDAP) can be used to enquire as to the status of a particular multicast group.

The hosts that show interest in secured groups MUST be operated by authenticated and authorized end users. One approach to the task of authentication and authorization of end users is based on the use of PANA [RFC5191] and EAP [RFC3748], and is described in [I-D.atwood-mboned-mrac-req], [I-D.atwood-mboned-mrac-arch] and draft-atwood-mboned-pana (not yet published).

A coordination protocol may be needed to manage and distribute the Security Associations (SAs) for secured groups among the routers and the hosts that correspond to authenticated and authorized end users. One set of possible procedures for SA creation and maintenance is specified in draft-atwood-pim-gsam (not yet published).

## 2. Overview of SIGMP

SIGMP is an extension to IGMP and performs receiver access control for groups in a multicast network. It retains most mechanisms of IGMP and has two working modes: 1) mode compatible with IGMP v2 and 2) mode compatible with IGMP v3. It works in either mode and is transparent for hosts that support IGMP, but not SIGMP. In addition, SIGMP uses IPsec to secure part of its packets. For an open group, it delivers the data to any host unconditionally as IGMP does. However, for a secured group, SIGMP only delivers the data to the hosts that have established SAs in the IPsec subsystem in order to perform access control.

In a network segment, hosts show their interest in secured groups using IPsec protected packets although their interest for open groups is still reported using unprotected packets. Similarly, routers query the membership interest for a secured group using IPsec protected packets, although the general query and the query for the membership of open groups are performed using unprotected packets.

In general, the packets in SIGMP are classified into four categories, which are Query for Open Group (OGQ), Query for Secured Group (SGQ), Report for Open Group (OGR) and Report for Secured Group (SGR). OGQ and SGQ are sent by the Querier and are used to learn the membership of open groups (or all groups for general query) and secured groups respectively. In detail, OGQ includes general query, specific-group query for open group and group-and-source-specific

query for the source of open group. SGQ includes specific-group query for secured group and group-and-source-specific query for the source of secured group. OGR and SGR are sent by hosts and used to report the membership of open groups and secured groups respectively. In detail, OGR includes report to specific-group query for open group, report to group-and-source-specific query for the source of open group, unsolicited report for open group and part of reports to general query. SGR includes report to specific-group query for secured group, report to group-and-source-specific query for the source of secured group, unsolicited report for secured group and part of reports to general query. SGQ and SGR are protected by IPsec at IP layer while OGQ and OGR are delivered without IPsec protection.

The destination address of packets in IP layer is specified as follows. In SGQ and SGR, the destination address is a secured group address. In OGQ, it is 224.0.0.1 if the packet is general query and otherwise it is an open group address. In OGR, it is 224.0.0.22 if the packet is the report to general query compatible with IGMP v3 and otherwise it is an open group address. The two addresses of 224.0.0.1 and 224.0.0.22 are the open group addresses. NOTE: When SIGMP works in the mode compatible with IGMP v3, the response to a general query contains zero or one OGR and zero or more SGR. It is described in detail in Section 5.2.1.

### 3. Packet Format

The packet format of SIGMP is identical to the packet format for IGMP. In detail, the format is the same as IGMP v2 when SIGMP works in the mode compatible with IGMP v2. The format is the same as IGMP v3 when SIGMP works in the mode compatible with IGMP v3.

### 4. Router Operations

Router operations in SIGMP are based on router operations in IGMP. However, some additional operations must be appended since access control to secured groups is extended into SIGMP. This section describes the additional operations for the two working modes.

#### 4.1. Router Operations Compatible with IGMP v2

The additional router operations focus on the operations for a received report.



#### 4.1.1. Router Operations for a Received Report

On receiving a report, a router checks the group address in the received report. If the group address indicates an open group, the report is considered as an OGR. A router will process an OGR as it does that in IGMP v2 directly. Otherwise, the received report is an SGR that SHOULD just have been authenticated (and decrypted) by the IPsec subsystem (e.g., AH [RFC4302]). For SGR, a router must perform two verifications: address consistency and SA existence.

In the address consistency verification, a router compares two addresses: the group address in the SIGMP report and the destination address in the IP header. The verification fails if the two addresses are not the same. In the failure case, the sender of the IGMP Report has attempted to hide a request for a specific group (probably a secured group) in an IGMP Report for a different group (probably an open group). This will cause the IPsec subsystem to deliver the IGMP Report without requiring it to be protected. Therefore a router must discard the report if this address consistency verification fails.

In the SA existence verification, a router checks whether SAs have been established for the secured group whose address is contained in the received report. The verification fails if there are no valid SAs for the group in the router's IPsec subsystem. Since the IPsec subsystem is used to enforce the access control, no access to a secured group is permitted until its SAs have been established. Therefore a router must discard the report if this verification fails.

If the two verifications succeed on SGR, a router will proceed to update the group memberships and refresh the timers as it does in IGMP v2. In summary, the router operations for a received report are shown in Table 1.

#	Group Address	Address Consistency	SA Existence	Operations for Report
1	Open	-	-	Process as IGMP v2
2	Secured	No	-	Discard
3	Secured	Yes	No	Discard
4	Secured	Yes	Yes	Process as IGMP v2

Table 1: Router Operations for a Received Report for the Mode Compatible with IGMP v2

#### 4.2. Router Operations Compatible with IGMP v3

The additional router operations still focus on the operations for a received report. However, there is a little difference between the operations in the mode compatible with IGMP v3 and the operations in the mode compatible with IGMP v2, since the formats of received reports in the two modes are different.

##### 4.2.1. Router Operations on a Received Report

On receiving a report, a router checks the number of group records in the report. If the number is more than one, it indicates that the report is an OGR, but not an SGR, since only one group record is included in an SGR. In this case, every group record in the report must be verified further as follows. A router checks the multicast address in the group record. If the multicast address is an open group address, a router will process the group record as it does in IGMP v3. Otherwise, a secured group address is in the group record and a router must discard the group record. The OGR including more than one group records is not protected by IPsec systems and is not permitted to contain any information related to any secured group.

In contrast, if the number of the group records is just one, a router still checks the multicast address in the single group record. If the multicast address indicates an open group address, the received report is considered as an OGR and a router will process the group record as it does that in IGMP v3 directly. Otherwise, the received report SHOULD be an SGR that SHOULD just be authenticated (and decrypted) by the IPsec subsystem. For the single group record in the SGR, a router must perform two verifications, address consistency and SA existence, similar to Section 4.1.

In the address consistency verification, a router compares two addresses: the multicast address in the group record of the SIGMP report and the destination address in the IP header. A router must discard the report if the two addresses are not the same.

In SA existence verification, a router checks whether SAs have been established for the secured group whose address is contained in the group record of the received report. A router must discard the report if there are no SAs established in the router's IPsec subsystem.

If the two verifications succeed on an SGR, a router will proceed to update the group memberships and refresh the timers as it does in IGMP v3. In summary, router operations for a received report are shown in Table 2.

#	#Group record in report	Multicast Address in Group Record	Address Consistency	SA Existence	Operations for Group Record
1	>1	Open	-	-	Process as IGMP v2
2	>1	Secured	-	-	Discard
3	=1	Open	-	-	Process as IGMP v2
4	=1	Secured	No	-	Discard
5	=1	Secured	Yes	No	Discard
6	=1	Secured	Yes	Yes	Process as IGMP v2

Table 2: Router Operations for a Received Report for Mode Compatible with IGMP v3

## 5. Host Operations

Host operations in SIGMP are based on host operations in IGMP. However, some additional operations must be appended since access control to secured group is extended into SIGMP. This section describes the additional operations for the two working modes.

### 5.1. Host Operations Compatible with IGMP v2

The additional host operations focus on the conditions for unsolicited report and the operations for a received query.

#### 5.1.1. Conditions for Unsolicited Report

Before creating an unsolicited report, a host must check the reported group. If the report group is open, a host will do as in IGMP v2. If secured, a host must continue to check whether SAs have been established for the secured group. If no SA is defined for this group address, a host MUST return an error indication to the issuer of the request that provoked the unsolicited report. [[Is this the right behavior?]]

#### 5.1.2. Host Operations for a Received Query

On receiving the query, a host does the additional operation as a router does in Section 4.2.1.

### 5.2. Host Operations Compatible with IGMP v3

The additional host operations focus on three aspects: 1) the conditions for unsolicited report, 2) the operations for a received non-general query and 3) the operations for a received general query. The first two are identical to those described in Section 5.1.1 and Section 5.1.2. In this subsection, only the last case is explained.

#### 5.2.1. Host Operations for a Received General Query

When it determines to respond to a general query, a host creates zero or one OGR and zero or more SGR in SIGMP instead of one report in IGMP v3. The OGR reports the current state of all the open groups that the host is interested in. Each SGR reports the current state of one secured group that the host is interested in.

At the IP layer, the destination address of OGR is 224.0.0.22. In contrast, at the IP layer the destination addresses of SGRs are the secured group addresses. Since IPsec has established SAs for secured groups, SGRs will be protected and the OGR will not.

### 6. IANA Considerations

The protocol number of SIGMP is the same as IGMP.

### 7. References

## 7.1. 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.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

## 7.2. Informative References

- [I-D.atwood-mboned-mrac-arch]  
william.atwood@concordia.ca, w., Li, B., and S. Islam,  
"Architecture for IP Multicast Receiver Access Control",  
draft-atwood-mboned-mrac-arch-00 (work in progress),  
October 2013.
- [I-D.atwood-mboned-mrac-req]  
william.atwood@concordia.ca, w., Islam, S., and B. Li,  
"Requirements for IP Multicast Receiver Access Control",  
draft-atwood-mboned-mrac-req-00 (work in progress),  
October 2013.
- [RFC3748] Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., and H. Levkowetz, "Extensible Authentication Protocol (EAP)", RFC 3748, June 2004.
- [RFC4302] Kent, S., "IP Authentication Header", RFC 4302, December 2005.
- [RFC5191] Forsberg, D., Ohba, Y., Patil, B., Tschofenig, H., and A. Yegin, "Protocol for Carrying Authentication for Network Access (PANA)", RFC 5191, May 2008.

## Authors' Addresses

William Atwood  
Concordia University/CSE  
1455 de Maisonneuve Blvd, West  
Montreal, QC H3G 1M8  
Canada

Phone: +1(514)848-2424 ext3046  
Email: [william.atwood@concordia.ca](mailto:william.atwood@concordia.ca)  
URI: <http://users.encs.concordia.ca/~bill>

Bing Li  
Concordia University/CSE  
1455 de Maisonneuve Blvd, West  
Montreal, QC H3G 1M8  
Canada

Email: [leebingice@gmail.com](mailto:leebingice@gmail.com)

Layer 2 Virtual Private Networks  
Internet-Draft  
Intended status: Informational  
Expires: April 11, 2014

O. Dornon  
J. Kotalwar  
Alcatel-Lucent  
V. Hemige

R. Qiu  
Z. Zhang  
Juniper Networks, Inc.  
October 8, 2013

PIM Snooping over VPLS  
draft-ietf-l2vpn-vpls-pim-snooping-05

Abstract

This document describes the procedures and recommendations for VPLS PEs to facilitate replication of multicast traffic to only certain ports (behind which there are interested PIM routers and/or IGMP hosts) via PIM Snooping and PIM Proxy.

With PIM Snooping, PEs passively listen to certain PIM control messages to build control and forwarding states while transparently flooding those messages. With PIM Proxy, PEs do not flood PIM Join/Prune messages but only generate their own and send out of certain ports, based on the control states built from downstream Join/Prune messages. PIM Proxy is required when PIM Join suppression is enabled on the CE devices and useful to reduce PIM control traffic in a VPLS domain.

The document also describes PIM Relay, which can be viewed as light-weight proxy, where all downstream Join/Prune messages are simply forwarded out of certain ports but not flooded to avoid triggering PIM Join suppression on CE devices.

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 April 11, 2014.

#### 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 . . . . .	5
1.1. Multicast Snooping in VPLS . . . . .	5
1.2. Assumptions . . . . .	6
1.3. Definitions . . . . .	7
2. PIM Snooping for VPLS . . . . .	7
2.1. PIM protocol background . . . . .	7
2.2. General Rules for PIM Snooping in VPLS . . . . .	8
2.2.1. Preserving Assert Trigger . . . . .	8
2.3. Some Considerations for PIM Snooping . . . . .	9
2.3.1. Scaling . . . . .	9
2.3.2. IPv6 . . . . .	10
2.3.3. PIM-SM (*,*,RP) . . . . .	10
2.4. PIM Snooping vs PIM Proxy . . . . .	10
2.4.1. Differences between PIM Snooping, Relay and Proxy . . . . .	10
2.4.2. PIM Control Message Latency . . . . .	11
2.4.3. When to Snoop and When to Proxy . . . . .	12
2.5. Discovering PIM Routers . . . . .	13
2.6. PIM-SM and PIM-SSM . . . . .	14
2.6.1. Building PIM-SM Snooping States . . . . .	14
2.6.2. Explanation for per (S,G,N) states . . . . .	17
2.6.3. Receiving (*,G) PIM-SM Join/Prune Messages . . . . .	17
2.6.4. Receiving (S,G) PIM-SM Join/Prune Messages . . . . .	19
2.6.5. Receiving (S,G,rpt) Join/Prune Messages . . . . .	21
2.6.6. Sending Join/Prune Messages Upstream . . . . .	21
2.7. Bidirectional-PIM (PIM-BIDIR) . . . . .	22
2.8. Interaction with IGMP Snooping . . . . .	23
2.9. PIM-DM . . . . .	23
2.9.1. Building PIM-DM Snooping States . . . . .	23
2.9.2. PIM-DM Downstream Per-Port PIM(S,G,N) State Machine . . . . .	24
2.9.3. Triggering ASSERT election in PIM-DM . . . . .	24
2.10. PIM Proxy . . . . .	24
2.10.1. Upstream PIM Proxy behavior . . . . .	24
2.11. Directly Connected Multicast Source . . . . .	25
2.12. Data Forwarding Rules . . . . .	25
2.12.1. PIM-SM Data Forwarding Rules . . . . .	26
2.12.2. PIM-DM Data Forwarding Rules . . . . .	27
3. IANA Considerations . . . . .	28
4. Security Considerations . . . . .	28
5. Contributors . . . . .	28
6. Acknowledgements . . . . .	29
7. References . . . . .	29
7.1. Normative References . . . . .	29
7.2. Informative References . . . . .	29
Appendix A. PIM-BIDIR Thoughts . . . . .	30
A.1. PIM-BIDIR Data Forwarding Rules . . . . .	30
Appendix B. Example Network Scenario . . . . .	31

B.1. Pim Snooping Example . . . . .	32
B.2. PIM Proxy Example with (S,G) / (*,G) interaction . . . . .	34
Authors' Addresses . . . . .	40

## 1. Introduction

In Virtual Private LAN Service (VPLS), the Provider Edge (PE) devices provide a logical interconnect such that Customer Edge (CE) devices belonging to a specific VPLS instance appear to be connected by a single LAN. Forwarding Information Base for a VPLS instance is populated dynamically by source MAC address learning. Once a unicast MAC address is learned and associated with a particular Attachment Circuit (AC) or PseudoWire (PW), a frame destined to that MAC address only needs to be sent on that AC or PW.

For a frame not addressed to a known unicast MAC address, flooding has to be used. This happens with the following so called BUM traffic:

- o B: The destination MAC address is a broadcast address,
- o U: The destination MAC address is unknown (has not been learned),
- o M: The destination MAC address is a multicast address.

Multicast frames are flooded because a PE cannot know where multicast members reside. VPLS solutions (i.e., [VPLS-LDP] and [VPLS-BGP]) perform replication for multicast traffic at the ingress PE devices. As stated in the VPLS Multicast Requirements draft [VPLS-MCAST-REQ], there are two issues with VPLS Multicast today:

- o A. Multicast traffic is replicated to non-member sites.
- o B. Replication of PWs on shared physical path.

Issue A can be solved by Multicast Snooping - PEs learn sites with multicast members by snooping multicast protocol control messages and forward IP multicast traffic only to member sites. This document describes the procedures to achieve that when PIM is running between the CE devices. Issue B is outside the scope of this document and discussed in [VPLS-MCAST-TREES].

While this document is in the context of VPLS, the procedures apply to regular layer-2 switches interconnected by physical connections as well. In that case, the PW related concept/procedures are not applicable and that's all.

### 1.1. Multicast Snooping in VPLS

IGMP Snooping procedures described in [IGMP-SNOOP] make sure that IP multicast traffic is only sent out of the following:

- o Attachment Circuits (ACs) connecting to hosts that report related group membership
- o ACs connecting to routers
- o PseudoWires (PWs) connecting to remote PEs that have the above described ACs

Notice that traffic is always sent out of ports connecting to routers, even those on which there are no snooped group memberships, because IGMP Snooping alone can not determine if there are interested receivers beyond those routers. To further restrict traffic sent to those routers, PIM Snooping can be used, and this document describes the procedures, including the rules when both IGMP and PIM are active in a VPLS instance.

Note that for both IGMP and PIM, the term Snooping is used loosely, referring to the fact that a layer-2 device peeks into layer-3 routing protocol messages to build relevant control and forwarding states. Depending on how the control messages are handled (transparently flooded, selectively forwarded, or consumed and then regenerated), the procedure/process may be called Snooping or Proxy in different contexts.

Unless explicitly noted, the procedures in this document are used for either PIM Snooping or PIM Proxy, and we will largely refer to PIM "Snooping" in this document. The PIM Proxy specific procedures are described in Section 2.6.6. Differences that need to be observed while implementing one or the other and recommendations on which method to employ in different scenarios are noted in section Section 2.4.

This document also describes PIM Relay, which can be viewed as light-weight Proxy. Unless explicitly noted, in the rest of the document Proxy implicitly includes Relay as well.

## 1.2. Assumptions

The document assumes that the reader has a good understanding of the PIM protocols. The text in this draft is written in the same style as the PIM RFCs to help correlate the concepts and to make it easier to follow. In order to avoid replicating the text relating to PIM protocol handling here, this draft cross references into definitions of macros and procedures from the PIM RFCs, and assumes that the user will infer such detail from those PIM RFCs. Deviations in protocol handling specific to PIM Snooping are specified in this draft.

### 1.3. Definitions

There are several definitions referenced in this document that are well described in the PIM RFCs [PIM-SM], PIM-BIDIR, PIM-DM]. The following definitions and abbreviations are used throughout this document:

- o A port is defined as either an attachment circuit (AC) or a Pseudo-Wire (PW).
- o When we say a PIM message is 'received' on a port, it means that a PIM Snooping PE snooped the PIM message.

Abbreviations used in the document:

- o S: IP Address of the Multicast Source.
- o G: IP Address of the Multicast Group.
- o N: Upstream Neighbor field in a Join/Prune/Graft message.
- o Rport(N): Port on which neighbor N is learnt.

Other definitions are explained in the sections where they are introduced.

## 2. PIM Snooping for VPLS

### 2.1. PIM protocol background

PIM is a multicast routing protocol running between routers, which are CE devices in a VPLS. PIM shares many of the common characteristics of a routing protocol, such as discovery messages (e.g., neighbor discovery using Hello messages), topology information (e.g., multicast tree), and error detection and notification (e.g., dead timer and designated router election). PIM does not participate in exchange of unicast routing databases, but it uses the unicast routing table to provide reverse path information for building multicast trees. There are a few variants of PIM. In [PIM-DM], multicast data is pushed towards the members similar to broadcast mechanism but routers without attached receivers will prune back towards the source. Unlike PIM-DM, other PIM flavors (PIM-SM [PIM-SM], PIM-SSM [PIM-SSM], and PIM-BIDIR [PIM-BIDIR]) employs a pull methodology via explicit joins instead of push technique.

PIM routers periodically exchange Hello messages to discover and maintain stateful sessions with neighbors. After neighbors are

discovered, PIM routers can signal their intentions to join or prune specific multicast groups. This is accomplished by having downstream routers send an explicit Join/Prune message (for the sake of generalization, consider Graft messages for PIM-DM as Join messages) to the upstream routers. The Join/Prune message can be group specific (\*,G) or group and source specific (S,G).

## 2.2. General Rules for PIM Snooping in VPLS

The following rules for the correct operation of PIM snooping MUST be followed.

- o PIM Snooping MUST NOT affect the operation of customer layer-2 protocols (e.g., BPDUs) or layer-3 protocols.
- o PIM messages and multicast data traffic forwarded by PEs MUST follow the split-horizon rule for mesh PWs.
- o PIM snooping states in a PE MUST be per VPLS instance.
- o PIM assert triggers MUST be preserved to the extent necessary to avoid sending duplicate traffic to the same PE (see Section 2.2.1).

### 2.2.1. Preserving Assert Trigger

In PIM-SM/DM, there are scenarios where multiple routers could be forwarding the same multicast traffic on a LAN. When this happens, using PIM Assert Election process by sending PIM Assert Messages, routers ensure that only the Assert Winner forwards traffic on the LAN. The Assert Election is a data driven event and happens only if a router sees traffic on the interface to which it should be forwarding the traffic. In the case of VPLS with snooping, two routers may forward the same flow at the same time but each copy may reach different set of PEs, and that is acceptable from the point of view of avoiding duplicate traffic. If the two copies may reach the same PE then the sending routers must be able to see each other's traffic, in order to trigger Assert Election and stop duplicate traffic.

To achieve that, PIM-SM Snooping MUST not only forward multicast traffic for an (S,G) on the ports on which they snooped Joins(S,G)/Joins(\*,G), but also towards the upstream neighbor(s)). In other words, the ports on which the upstream neighbors are learnt must be added to the outgoing port list along with the ports on which Joins are snooped.

Similarly, PIM-DM Snooping SHOULD make sure that asserts can be

triggered (Section 2.9.3).

The above logic needs to be facilitated without breaking VPLS Split Horizon Rules. i.e. traffic should not be forwarded on the port on which it was received, and traffic arriving on a PW MUST NOT be forwarded onto other PW(s).

### 2.3. Some Considerations for PIM Snooping

The PIM Snooping solution described here requires a PE to examine and operate on only PIM Hello and PIM Join/Prune packets. The PE does not need to examine any other PIM packets.

Most of the procedures in PIM Snooping in the handling of PIM Hellos and PIM Join/Prune packets are very similar to that of a PIM Router.

However, the PE does not need to have any routing tables like is required in PIM Multicast Routing. It knows how to forward Join/Prunes by looking at the Upstream Neighbor field in the Join/Prune packets.

The PE does not need to know about Rendezvous Points (RP) and does not have to maintain any RP Set. All that is transparent to a PIM Snooping PE.

In the following sub-sections, we list some considerations and observations for the implementation of PIM Snooping in VPLS.

#### 2.3.1. Scaling

Snooping needs to be employed on ACs at the downstream PEs to prevent traffic from being sent out of ACs unnecessarily. Snooping techniques can also be employed on PWs at the upstream PEs to prevent traffic from being sent to PEs unnecessarily. This may work well for small to medium scale deployments. However, if there are a large number of VPLS instances with a large number of PEs per instances, then the amount of snooping required at the upstream PEs can overwhelm the upstream PEs.

There are two methods to reduce the burden on the upstream PEs. One is to use PIM Proxy as described in Section 2.6.6, to reduce the control messages forwarded by a PE. The other is not to snoop on the PWs at all, but PEs signal the snooped states to other PEs out of band via BGP, as described in [VPLS-MCAST-TREES]. In this document, it is assumed that Snooping is performed on PWs.

### 2.3.2. IPv6

In VPLS, PEs forward Ethernet frames received from CEs and as such are agnostic of the layer-3 protocol used by the CEs. However, as an IGMP and PIM snooping PE, the PE would have to look deeper into the IP and IGMP/PIM packets and build snooping state based on that. The PIM Protocol specifications handle both IPv4 and IPv6. The specification for PIM Snooping in this draft can be applied to both IPv4 and IPv6 payloads.

### 2.3.3. PIM-SM (\*,\*,RP)

This draft does not address (\*,\*,RP) states in the VPLS network. Although [PIM-SM] specifies that routers MUST support (\*,\*,RP) states, there are very few implementations that actually support it in actual deployments, and it is being removed from the PIM protocol in its ongoing advancement process in IETF. Given that, this draft omits the specification relating to (\*,\*,RP) support.

## 2.4. PIM Snooping vs PIM Proxy

The document has previously alluded to PIM Snooping/Relay/Proxy. Details on the PIM Proxy/Relay solution are discussed in Section 2.6.6. In this section, a brief description and comparison are given.

### 2.4.1. Differences between PIM Snooping, Relay and Proxy

Differences between PIM Snooping and Proxy/Relay can be summarized as the following:

PIM Snooping	PIM Relay	PIM Proxy
Join/Prune messages snooped and flooded everywhere	Join/Prune messages snooped; forwarded as is out of certain upstream ports	Join/Prune messages consumed. Regenerated ones sent out of certain upstream ports
No PIM packets generated.	No PIM packets generated	New Join/Prune messages generated
CE Join Suppression not allowed	CE Join Suppression allowed	CE Join Suppression allowed

Note that the differences apply only to PIM Join/Prune messages. PIM



Hello messages are snooped and flooded in all cases.

Other than the above differences, most of the procedures are common to PIM Snooping and PIM Proxy/Relay, unless specifically stated otherwise.

Pure PIM Snooping PEs simply snoop on PIM packets as they are being forwarded in the VPLS. As such they truly provide transparent LAN services since no customer packets are modified or consumed or new packets introduced in the VPLS. It is also simpler to implement than PIM Proxy. However for PIM Snooping to work correctly, it is a requirement that CE routers MUST disable Join suppression in the VPLS.

Given that a large number of existing CE deployments do not support disabling of Join suppression and given the operational complexity for a provider to manage disabling of Join suppression in the VPLS, it becomes a difficult solution to deploy. Another disadvantage of PIM Snooping is that it does not scale as well as PIM Proxy. If there are a large number of CEs in a VPLS, then every CE will see every other CE's Join/Prune messages.

PIM Proxy/Relay has the advantage that it does not require Join suppression to be disabled in the VPLS. Multicast as a VPLS service can be very easily provided without requiring any changes on the CE routers. PIM Proxy/Relay helps scale VPLS Multicast since Join/Prune messages are only sent to certain upstream ports instead of flooded, and in case of full Proxy (vs. Relay) the PEs intelligently generate only one Join/Prune message for a given flow.

PIM Proxy however loses the transparency argument since Join/Prunes could get modified or even consumed at a PE. Also, new packets could get introduced in the VPLS. However, this loss of transparency is limited to PIM Join/Prune packets. It is in the interest of optimizing multicast in the VPLS and helping a VPLS network scale much better. Data traffic will still be completely transparent.

#### 2.4.2. PIM Control Message Latency

A PIM Snooping/Proxy/Relay PE snoops on PIM Hello packets while transparently flooding them in the VPLS. As such there is no latency introduced by the VPLS in the delivery of PIM Hello packets to remote CEs in the VPLS.

A PIM Snooping PE snoops on PIM Join/Prune packets while transparently flooding them in the VPLS. There is no latency introduced by the VPLS in the delivery of PIM Join/Prune packets when PIM Snooping is employed.

A PIM Proxy/Relay PE does not simply flood PIM Join/Prune packets. This can result in additional latency for a downstream CE to receive multicast traffic after it has sent a Join. When a downstream CE prunes a multicast stream, the traffic should stop flowing to the CE with no additional latency introduced by the VPLS.

Performing only proxy of Join/Prune and not Hello messages keeps the PE behavior very similar to that of a PIM router without introducing too much additional complexity. It keeps the PIM Proxy solution fairly simple. Since Join/Prunes are forwarded by a PE along the slow-path and all other PIM packet types are forwarded along the fast-path, it is very likely that packets forwarded along the fast-path will arrive "ahead" of Join/Prune packets at a CE router (note the stress on the fact that fast-path messages will never arrive after Join/Prunes). Of particular importance are Hello packets sent along the fast-path. We can construct a variety of scenarios resulting in out of order delivery of Hellos and Join/Prune messages. However, there should be no deviation from normal expected behavior observed at the CE router receiving these messages out of order.

#### 2.4.3. When to Snoop and When to Proxy

From the above descriptions, factors that affect the choice of Snooping/Relay/Proxy include:

- o Whether CEs do Join Suppression or not
- o Whether Join/Prune latency is critical or not
- o Whether the scale of PIM protocol message/states in a VPLS requires the scaling benefit of Proxy

Of the above factors, Join Suppression is the hard one - pure Snooping can only be used when Join Suppression is disabled on all CEs. The latency associated with Relay/Proxy is implementation dependent and may not be a concern at all with a particular implementation. The scaling benefit may not be important either, in that on a real LAN with Explicit Tracking (ET) a PIM router will need to receive and process all PIM Join/Prune messages as well.

A PIM router indicates that Join Suppression is disabled if the T-bit is set in the LAN Prune Delay option of its Hello message. If all PIM routers on a LAN set the T-bit, Explicit Tracking is possible, allowing an upstream router to track all the downstream neighbors that have Join states for any (S,G) or (\*,G). That has two benefits:

- o No need for PrunePending process - the upstream router may immediately stop forwarding data when it receives a Prune from the last downstream neighbor, and immediately prune to its upstream if that's for the last downstream interface.
- o For management purpose, the upstream router knows exactly which downstream routers exist for a particular Join State.

While full Proxy can be used with or without Join Suppression on CEs and does not interfere with an upstream CE's bypass of PrunePending process, it does proxy all its downstream CEs as a single one to the upstream, removing the second benefit mentioned above.

Therefore, the general rule is that if Join Suppression is enabled on CEs then Proxy or Relay MUST be used and if Suppression is known to be disabled on all CEs then either Snooping, Relay, or Proxy MAY be used while Snooping or Relay SHOULD be used.

An implementation MAY choose dynamic determination of which mode to use, through the tracking of the above mentioned T-bit in all snooped PIM Hello messages, or MAY simply require static provisioning.

## 2.5. Discovering PIM Routers

A PIM Snooping PE MUST snoop on PIM Hellos received on ACs and PWs. i.e. the PE transparently floods the PIM Hello while snooping on it. PIM Hellos are used by the snooping PE to discover PIM routers and their characteristics.

For each neighbor discovered by a PE, it includes an entry in the PIM Neighbor Database with the following fields:

- o Layer 2 encapsulation for the Router sending the PIM Hello.
- o IP Address and address family of the Router sending the PIM Hello.
- o Port (AC / PW) on which the PIM Hello was received.
- o Hello TLVs

The PE should be able to interpret and act on Hello TLVs currently defined in the PIM RFCs. The TLVs of particular interest in this document are:

- o Hello-Hold-Time
- o Tracking Support

- o DR Priority

Please refer to [PIM-SM] for a list of the Hello TLVs. When a PIM Hello is received, the PE MUST reset the neighbor-expiry-timer to Hello-Hold-Time. If a PE does not receive a Hello message from a router within Hello-Hold-Time, the PE MUST remove that neighbor from its PIM Neighbor Database. If a PE receives a Hello message from a router with Hello-Hold-Time value set to zero, the PE MUST remove that router from the PIM snooping state immediately.

From the PIM Neighbor Database, a PE MUST be able to use the procedures defined in [PIM-SM] to identify the PIM Designated Router in the VPLS instance. It should also be able to determine if Tracking Support is active in the VPLS instance.

## 2.6. PIM-SM and PIM-SSM

The key characteristic of PIM-SM and PIM-SSM is explicit join behavior. In this model, multicast traffic is only forwarded to locations that specifically request it. The root node of a tree is the Rendezvous Point (RP) in case of a shared tree (PIM-SM only) or the first hop router that is directly connected to the multicast source in the case of a shortest path tree. All the procedures described in this section apply to both PIM-SM and PIM-SSM, except for the fact that there is no (\*,G) state in PIM-SSM.

### 2.6.1. Building PIM-SM Snooping States

PIM-SM and PIM-SSM Snooping states are built by snooping on the PIM-SM Join/Prune messages received on AC/PWs.

The downstream state machine of a PIM-SM snooping PE very closely resembles the downstream state machine of PIM-SM routers. The downstream state consists of:

Per downstream (Port, \*, G):

- o DownstreamJPState: One of { "NoInfo" (NI), "Join" (J), "Prune Pending" (PP) }

Per downstream (Port, \*, G, N):

- o Prune Pending Timer (PPT(N))
- o Join Expiry Timer (ET(N))

Per downstream (Port, S, G):

- o DownstreamJPState: One of { "NoInfo" (NI), "Join" (J), "Prune Pending" (PP) }

Per downstream (Port, S, G, N):

- o Prune Pending Timer (PPT(N))
- o Join Expiry Timer (ET(N))

Per downstream (Port, S, G, rpt):

- o DownstreamJPRptState: One of { "NoInfo" (NI), "Pruned" (P), "Prune Pending" (PP) }

Per downstream (Port, S, G, rpt, N):

- o Prune Pending Timer (PPT(N))
- o Join Expiry Timer (ET(N))

Where S is the address of the multicast source, G is the Group address and N is the upstream neighbor field in the Join/Prune message. Notice that unlike on PIM-SM routers where PPT and ET are per (Interface, S, G), PIM Snooping PEs have to maintain PPT and ET per (Port, S, G, N). The reasons for this are explained in Section 2.6.2.

Apart from the above states, we define the following state summarization macros.

UpstreamNeighbors(\*,G): If there is one or more Join(\*,G) received on any port with upstream neighbor N and ET(N) is active, then N is added to UpstreamNeighbors(\*,G). This set is used to determine if a Join(\*,G) or a Prune(\*,G) with upstream neighbor N needs to be sent upstream.

UpstreamNeighbors(S,G): If there is one or more Join(S,G) received on any port with upstream neighbor N and ET(N) is active, then N is added to UpstreamNeighbors(S,G). This set is used to determine if a Join(S,G) or a Prune(S,G) with upstream neighbor N needs to be sent upstream.

UpstreamPorts(\*,G): This is the set of all Rport(N) ports where N is in the set UpstreamNeighbors(\*,G). Multicast Streams forwarded using a (\*,G) match MUST be forwarded to these ports in addition to downstream ports. So UpstreamPorts(\*,G) MUST be added to OutgoingPortList(\*,G).

UpstreamPorts(S,G): This is the set of all Rport(N) ports where N is in the set UpstreamNeighbors(S,G). UpstreamPorts(S,G) MUST be added to OutgoingPortList(S,G).

InheritedUpstreamPorts(S,G): This is the union of UpstreamPorts(S,G) and UpstreamPorts(\*,G).

UpstreamPorts(S,G,rpt): If PruneDesired(S,G,rpt) becomes true, then this set is set to UpstreamPorts(\*,G). Otherwise, this set is empty. UpstreamPorts(\*,G) (-) UpstreamPorts(S,G,rpt) MUST be added to OutgoingPortList(S,G).

UpstreamPorts(G): This set is the union of all the UpstreamPorts(S,G) and UpstreamPorts(\*,G) for a given G. Proxy (S,G) Join/Prune and (\*,G) Join/Prune messages MUST be sent to a subset of UpstreamPorts(G) as specified in Section 2.6.6.1.

PWPorts: This is the set of all PWs.

OutgoingPortList(\*,G): This is the set of all ports to which traffic needs to be forwarded on a (\*,G) match.

OutgoingPortList(S,G): This is the set of all ports to which traffic needs to be forwarded on an (S,G) match.

See Section 2.12 on Data Forwarding Rules for the specification on how OutgoingPortList is calculated.

NumETsActive(Port,\*,G): Number of (Port,\*,G,N) entries that have Expiry Timer running. This macro keeps track of the number of Join(\*,G)s that are received on this Port with different upstream neighbors.

NumETsActive(Port,S,G): Number of (Port,S,G,N) entries that have Expiry Timer running. This macro keeps track of the number of Join(S,G)s that are received on this Port with different upstream neighbors.

RpfVectorTlvs(\*,G): RPF Vectors [RPF-VECTOR] are TLVs that may be present in received Join(\*,G) messages. If present, they must be copied to RpfVectorTlvs(\*,G).

RpfVectorTlvs(S,G): RPF Vectors [RPF-VECTOR] are TLVs that may be present in received Join(S,G) messages. If present, they must be copied to RpfVectorTlvs(S,G).

Since there are a few differences between the downstream state machines of PIM-SM Routers and PIM-SM snooping PEs, we specify the

details of the downstream state machine of PIM-SM snooping PEs at the risk of repeating most of the text documented in [PIM-SM].

#### 2.6.2. Explanation for per (S,G,N) states

In PIM Routing protocols, states are built per (S,G). On a router, an (S,G) has only one RPF-Neighbor. However, a PIM Snooping PE does not have the Layer 3 routing information available to the routers in order to determine the RPF-Neighbor for a multicast flow. It merely discovers it by snooping the Join/Prune message. A PE could have snooped on two or more different Join/Prune messages for the same (S,G) that could have carried different Upstream-Neighbor fields. This could happen during transient network conditions or due to dual-homed sources. A PE cannot make assumptions on which one to pick, but instead must facilitate the CE routers decide which Upstream Neighbor gets elected the RPF-Neighbor. And for this purpose, the PE will have to track downstream and upstream Join/Prune per (S,G,N).

#### 2.6.3. Receiving (\*,G) PIM-SM Join/Prune Messages

A Join(\*,G) or Prune(\*,G) is considered "received" if the following conditions are met:

- o The port on which it arrived is not Rport(N) where N is the upstream-neighbor N of the Join/Prune(\*,G), or,
- o if both RPort(N) and the arrival port are PWs, then there exists at least one other (\*,G,Nx) or (Sx,G,Nx) state with an AC UpstreamPort.

For simplicity, the case where both RPort(N) and the arrival port are PWs is referred to as PW-only Join/Prune in this document. The PW-only Join/Prune handling is so that the RPort(N) PW can be added to the related forwarding entries' OutgoingPortList to trigger Assert, but that is only needed for those states with AC UpstreamPort. Note that in PW-only case, it is OK for the arrival port and RPort(N) to be the same. See Appendix Appendix B for examples.

When a router receives a Join(\*,G) or a Prune(\*,G) with upstream neighbor N, it must process the message as defined in the state machine below. Note that the macro computations of the various macros resulting from this state machine transition is exactly as specified in the PIM-SM RFC [PIM-SM].

We define the following per-port (\*,G,N) macro to help with the state machine below.

Figure 1 : Downstream per-port (\*,G) state machine in tabular form

Event	Previous State		
	NoInfo (NI)	Join (J)	Prune-Pend
Receive Join(*,G)	-> J state Action RxJoin(N)	-> J state Action RxJoin(N)	-> J state Action RxJoin(N)
Receive Prune(*,G) and NumETsActive<=1	-	-> PP state Start PPT(N)	-> PP state
Receive Prune(*,G) and NumETsActive>1	-	-> J state Start PPT(N)	-
PPT(N) expires	-	-> J state Action PPTExpiry(N)	-> NI state Action PPTExpiry(N)
ET(N) expires and NumETsActive<=1	-	-> NI state Action ETExpiry(N)	-> NI state Action ETExpiry(N)
ET(N) expires and NumETsActive>1	-	-> J state Action ETExpiry(N)	-

Action RxJoin(N):

If ET(N) is not already running, then start ET(N). Otherwise restart ET(N). If N is not already in UpstreamNeighbors(\*,G), then add N to UpstreamNeighbors(\*,G) and trigger a Join(\*,G) with upstream neighbor N to be forwarded upstream. If there are RPF Vector TLVs in the received (\*,G) message and if they are different from the recorded RpfVectorTlvs(\*,G), then copy them into RpfVectorTlvs(\*,G).

Action PPTExpiry(N):

Same as Action ETExpiry(N) below, plus Send a Prune-Echo(\*,G) with upstream-neighbor N on the downstream port.

Action ETExpiry(N):



Disable timers ET(N) and PPT(N). Delete neighbor state (Port,\*,G,N). If there are no other (Port,\*,G) states with NumETsActive(Port,\*,G) > 0, transition DownstreamJPState to NoInfo. If there are no other (Port,\*,G,N) state (different ports but for the same N), remove N from UpstreamPorts(\*,G) - this also serves as a trigger for US FSM (JoinDesired(\*,G,N) becomes FALSE).

#### 2.6.4. Receiving (S,G) PIM-SM Join/Prune Messages

A Join(S,G) or Prune(S,G) is considered "received" if the following conditions are met:

- o The port on which it arrived is not Rport(N) where N is the upstream-neighbor N of the Join/Prune(S,G), or,
- o if both RPort(N) and the arrival port are PWs, then there exists at least one other (\*,G,Nx) or (S,G,Nx) state with an AC UpstreamPort.

For simplicity, the case where both RPort(N) and the arrival port are PWs is referred to as PW-only Join/Prune in this document. The PW-only Join/Prune handling is so that the RPort(N) PW can be added to the related forwarding entries' OutgoingPortList to trigger Assert, but that is only needed for those states with AC UpstreamPort. See Appendix Appendix B for examples.

When a router receives a Join(S,G) or a Prune(S,G) with upstream neighbor N, it must process the message as defined in the state machine below. Note that the macro computations of the various macros resulting from this state machine transition is exactly as specified in the PIM-SM RFC [PIM-SM].

Figure 2: Downstream per-port (S,G) state machine in tabular form

Event	Previous State		
	NoInfo (NI)	Join (J)	Prune-Pend
Receive Join(S,G)	-> J state Action RxJoin(N)	-> J state Action RxJoin(N)	-> J state Action RxJoin(N)
Receive Prune (S,G) and NumETsActive<=1	-	-> PP state Start PPT(N)	-> PP state
Receive Prune(S,G) and NumETsActive>1	-	-> J state Start PPT(N)	-
PPT(N) expires	-	-> J state Action PPTEpiry(N)	-> NI state Action PPTEpiry(N)
ET(N) expires and NumETsActive<=1	-	-> NI state Action ETExpiry(N)	-> NI state Action ETExpiry(N)
ET(N) expires and NumETsActive>1	-	-> J state Action ETExpiry(N)	-

**Action RxJoin(N):**

If ET(N) is not already running, then start ET(N). Otherwise, restart ET(N).

If N is not already in UpstreamNeighbors(S,G), then add N to UpstreamNeighbors(S,G) and trigger a Join(S,G) with upstream neighbor N to be forwarded upstream. If there are RPF Vector TLVs in the received (S,G) message and if they are different from the recorded RpfVectorTlvs(S,G), then copy them into RpfVectorTlvs(S,G).

**Action PPTEpiry(N):**

Same as Action ETExpiry(N) below, plus Send a Prune-Echo(S,G) with upstream-neighbor N on the downstream port.

**Action ETEpiry(N):**

Disable timers ET(N) and PPT(N). Delete neighbor state (Port,S,G,N). If there are no other (Port,S,G) states with NumETsActive(Port,S,G) > 0, transition DownstreamJPState to NoInfo. If there are no other (Port,S,G,N) state (different ports but for the same N), remove N from UpstreamPorts(S,G) - this also serves as a trigger for US FSM (JoinDesired(S,G,N) becomes FALSE).

**2.6.5. Receiving (S,G,rpt) Join/Prune Messages**

A Join(S,G,rpt) or Prune(S,G,rpt) is "received" when the port on which it was received is not also the port on which the upstream-neighbor N of the Join/Prune(S,G,rpt) was learnt.

While it is important to ensure that the (S,G) and (\*,G) state machines allow for handling per (S,G,N) states, it is not as important for (S,G,rpt) states. It suffices to say that the downstream (S,G,rpt) state machine is the same as what is defined in section 4.5.4 of the PIM-SM RFC [PIM-SM].

**2.6.6. Sending Join/Prune Messages Upstream**

This section applies only to a PIM Proxy/Relay PE and not to a PIM Snooping PE.

A full PIM Proxy (not Relay) PE MUST implement the Upstream FSM for which the procedures are similar to what is defined in section 4.5.6 of [PIM-SM].

For the purposes of the Upstream FSM, a Join or Prune message with upstream neighbor N is "seen" on a PIM Snooping PE if the port on which the message was received is also Rport(N), and the port is an AC. The AC requirement is needed because a Join received on the Rport(N) PW must not suppress this PE's Join on that PW.

A PIM Relay PE does not implement the Upstream FSM. It simply forwards received Join/Prune messages out of the same set of upstream ports as in the PIM Proxy case.

In order to correctly facilitate assert among the CE routers, such Join/Prunes need to sent not only towards the upstream neighbor, but also on certain PWs as described below.

If RpfVectorTlvs(\*,G) is not empty, then it must be encoded in a Join(\*,G) message sent upstream.

If RpfVectorTlvs(S,G) is not empty, then it must be encoded in a

Join(S,G) message sent upstream.

#### 2.6.6.1. Where to send Join/Prune messages

The following rules apply, to both forwarded (in case of PIM Relay), refresh and triggered (in case of PIM Proxy) (S,G)/(\*,G) Join/Prune messages.

- o The upstream neighbor field in the Join/Prune to be sent is set to the N in the corresponding Upstream FSM.
- o if Rport(N) is an AC, send the message to Rport(N).
- o Additionally, if OutgoingPortList(X,G,N) contains at least one AC, then the message MUST be sent to at least all the PWs in UpstreamPorts(G) (for (\*,G)) or InheritedUpstreamPorts(S,G) (for (S,G)). Alternatively, the message MAY be sent to all PWs.

Sending to a subset of PWs as described above guarantees that if traffic (of the same flow) from two upstream routers were to reach this PE, then the two routers will receive from each other, triggering assert.

Sending to all PWs guarantees that if two upstream routers both send traffic for the same flow (even if it is to different sets of downstream PEs), then they'll receive from each other, triggering assert.

#### 2.7. Bidirectional-PIM (PIM-BIDIR)

PIM-BIDIR is a variation of PIM-SM. The main differences between PIM-SM and Bidirectional-PIM are as follows:

- o There are no source-based trees, and source-specific multicast is not supported (i.e., no (S,G) states) in PIM-BIDIR.
- o Multicast traffic can flow up the shared tree in PIM-BIDIR.
- o To avoid forwarding loops, one router on each link is elected as the Designated Forwarder (DF) for each RP in PIM-BIDIR.

The main advantage of PIM-BIDIR is that it scales well for many-to-many applications. However, the lack of source-based trees means that multicast traffic is forced to remain on the shared tree.

As described in [PIM-BIDIR], parts of a PIM-BIDIR enabled network may forward traffic without exchanging Join/Prune messages, for instance between DF's and the RPL.

As the described procedures for Pim snooping rely on the presence of Join/Prune messages, enabling Pim snooping on PIM-BIDIR networks could break the PIM-BIDIR functionality. Deploying Pim snooping on PIM-BIDIR enabled networks will require some further study. Some thoughts are gathered in Appendix A.

## 2.8. Interaction with IGMP Snooping

Whenever IGMP Snooping is enabled in conjunction with PIM Snooping in the same VPLS instance the PE SHOULD follow these rules:

- o To maintain the list of multicast routers and ports on which they are attached, the PE SHOULD NOT use the rules as described in RFC4541 [IGMP-SNOOP] but SHOULD rely on the neighbors discovered by PIM Snooping . This list SHOULD then be used to apply the forwarding rule as described in 2.1.1.(1) of RFC4541 [IGMP-SNOOP].
- o If the PE supports proxy-reporting, an IGMP membership learned only on a port to which a PIM neighbor is attached but not elsewhere SHOULD NOT be included in the summarized upstream report sent to that port.

## 2.9. PIM-DM

The characteristics of PIM-DM is flood and prune behavior. Shortest path trees are built as a multicast source starts transmitting.

### 2.9.1. Building PIM-DM Snooping States

PIM-DM Snooping states are built by snooping on the PIM-DM Join, Prune, Graft and State Refresh messages received on AC/PWs and State-Refresh Messages sent on AC/PWs. By snooping on these PIM-DM messages, a PE builds the following states per (S,G,N) where S is the address of the multicast source, G is the Group address and N is the upstream neighbor to which Prunes/Grafts are sent by downstream CEs:

Per PIM (S,G,N):

Port PIM (S,G,N) Prune State:

- \* DownstreamPState(S,G,N,Port): One of {"NoInfo" (NI), "Pruned" (P), "PrunePending" (PP)}
- \* Prune Pending Timer (PPT)

- \* Prune Timer (PT)
- \* Upstream Port (valid if the PIM(S,G,N) Prune State is "Pruned").

#### 2.9.2. PIM-DM Downstream Per-Port PIM(S,G,N) State Machine

The downstream per-port PIM(S,G,N) state machine is as defined in section 4.4.2 of [PIM-DM] with a few changes relevant to PIM Snooping. When reading section 4.4.2 of [PIM-DM] for the purposes of PIM-Snooping please be aware that the downstream states are built per (S, G, N, Downstream-Port) in PIM-Snooping and not per {Downstream-Interface, S, G} as in a PIM-DM router. As noted in the previous Section 2.9.1, the states (DownstreamPState) and timers (PPT and PT) are per (S,G,N,P).

#### 2.9.3. Triggering ASSERT election in PIM-DM

Since PIM-DM is a flood-and-prune protocol, traffic is flooded to all routers unless explicitly pruned. Since PIM-DM routers do not prune on non-RPF interfaces, PEs should typically not receive Prunes on Rport(RPF-neighbor). So the asserting routers should typically be in pim\_oiflist(S,G). In most cases, assert election should occur naturally without any special handling since data traffic will be forwarded to the asserting routers.

However, there are some scenarios where a prune might be received on a port which is also an upstream port (UP). If we prune the port from pim\_oiflist(S,G), then it would not be possible for the asserting routers to determine if traffic arrived on their downstream port. This can be fixed by adding pim\_iifs(S,G) to pim\_oiflist(S,G) so that data traffic flows to the UP ports.

#### 2.10. PIM Proxy

As noted earlier, PIM Snooping will work correctly only if Join Suppression is disabled in the VPLS. If Join Suppression is enabled in the VPLS, then PEs MUST do PIM Proxy/Relay for VPLS Multicast to work correctly. This section applies specifically to the full Proxy case and not Relay.

##### 2.10.1. Upstream PIM Proxy behavior

A PIM Proxy PE consumes Join/Prune messages and regenerates PIM Join/Prune messages to be sent upstream by implementing Upstream FSM as specified in the PIM RFC. This is the only difference from PIM Relay.

The source IP address in PIM packets sent upstream SHOULD be the address of a PIM downstream neighbor in the corresponding join/prune state. The address picked MUST NOT be the upstream neighbor field to be encoded in the packet. The layer 2 encapsulation for the selected source IP address MUST be the encapsulation recorded in the PIM Neighbor database for that IP address.

#### 2.11. Directly Connected Multicast Source

If there is a source in the CE network that connects directly into the VPLS instance, then multicast traffic from that source MUST be sent to all PIM routers on the VPLS instance apart from the IGMP receivers in the VPLS. If there is already (S,G) or (\*,G) snooping state that is formed on any PE, this will not happen per the current forwarding rules and guidelines. So, in order to determine if traffic needs to be flooded to all routers, a PE must be able to determine if the traffic came from a host on that LAN. There are three ways to address this problem:

- o The PE would have to do ARP snooping to determine if a source is directly connected.
- o Another option is to have configuration on all PEs to say there are CE sources that are directly connected to the VPLS instance and disallow snooping for the groups for which the source is going to send traffic. This way traffic from that source to those groups will always be flooded within the provider network.
- o A third option is to require that sources of CE multicast traffic must be behind a router.

This document recommends the third option - sources traffic must be behind a router.

#### 2.12. Data Forwarding Rules

First we define the rules that are common to PIM-SM and PIM-DM PEs. Forwarding rules for each protocol type is specified in the sub-sections.

If there is no matching forwarding state, then the PE SHOULD discard the packet, i.e., the UserDefinedPortList below SHOULD be empty.

The following general rules MUST be followed when forwarding multicast traffic in a VPLS:

- o Traffic arriving on a port MUST NOT be forwarded back onto the same port.
- o Due to VPLS Split-Horizon rules, traffic ingressing on a PW MUST NOT be forwarded to any other PW.

#### 2.12.1. PIM-SM Data Forwarding Rules

Per the rules in [PIM-SM] and per the additional rules specified in this document,

```
OutgoingPortList(*,G) = immediate_olist(*,G) (+)
                        UpstreamPorts(*,G) (+)
                        Rport(PimDR)
```

```
OutgoingPortList(S,G) = inherited_olist(S,G) (+)
                        UpstreamPorts(S,G) (+)
                        (UpstreamPorts(*,G) (-)
                        UpstreamPorts(S,G,rpt)) (+)
                        Rport(PimDR)
```

[PIM-SM] specifies how `immediate_olist(*,G)` and `inherited_olist(S,G)` are built. `PimDR` is the IP address of the PIM DR in the VPLS.

The PIM-SM Snooping forwarding rules are defined below in pseudocode:



```
BEGIN
  iif is the incoming port of the multicast packet.
  S is the Source IP Address of the multicast packet.
  G is the Destination IP Address of the multicast packet.

  If there is (S,G) state on the PE
  Then
    OutgoingPortList = OutgoingPortList(S,G)
  Else if there is (*,G) state on the PE
  Then
    OutgoingPortList = OutgoingPortList(*,G)
  Else
    OutgoingPortList = UserDefinedPortList
  Endif

  If iif is an AC
  Then
    OutgoingPortList = OutgoingPortList (-) iif
  Else
    ## iif is a PW
    OutgoingPortList = OutgoingPortList (-) PWPorts
  Endif

  Forward the packet to OutgoingPortList.
END
```

First if there is (S,G) state on the PE, then the set of outgoing ports is OutgoingPortList(S,G).

Otherwise if there is (\*,G) state on the PE, the set of outgoing ports is OutgoingPortList(\*,G).

The packet is forwarded to the selected set of outgoing ports while observing the general rules above in Section 2.12

#### 2.12.2. PIM-DM Data Forwarding Rules

The PIM-DM Snooping data forwarding rules are defined below in pseudocode:

```
BEGIN
    iif is the incoming port of the multicast packet.
    S is the Source IP Address of the multicast packet.
    G is the Destination IP Address of the multicast packet.

    If there is (S,G) state on the PE
    Then
        OutgoingPortList = olist(S,G)
    Else
        OutgoingPortList = UserDefinedPortList
    Endif

    If iif is an AC
    Then
        OutgoingPortList = OutgoingPortList (-) iif
    Else
        ## iif is a PW
        OutgoingPortList = OutgoingPortList (-) PWPorts
    Endif

    Forward the packet to OutgoingPortList.
END
```

If there is forwarding state for (S,G), then forward the packet to olist(S,G) while observing the general rules above in section Section 2.12

[PIM-DM] specifies how olist(S,G) is constructed.

### 3. IANA Considerations

This document makes no request of IANA.

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

### 4. Security Considerations

Security considerations provided in VPLS solution documents (i.e., [VPLS-LDP] and [VPLS-BGP]) apply to this document as well.

### 5. Contributors

Yetik Serbest, Suresh Boddapati co-authored earlier versions.

Karl (Xiangrong) Cai and Princy Elizabeth made significant contributions to bring the specification to its current state, especially in the area of Join forwarding rules.

## 6. Acknowledgements

Many members of the L2VPN and PIM working groups have contributed to and provided valuable comments and feedback to this draft, including Vach Kompella, Shane Amante, Sunil Khandekar, Rob Nath, Marc Lassere, Yuji Kamite, Yiqun Cai, Ali Sajassi, Jozef Raets, Himanshu Shah (Ciena), Himanshu Shah (Alcatel-Lucent).

## 7. References

### 7.1. Normative References

- [PIM-BIDIR] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, 2007.
- [PIM-DM] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast Version 2 - Dense Mode Specification", RFC 3973, 2005.
- [PIM-SM] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast- Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, 2006.
- [PIM-SSM] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, 2006.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, 1997.
- [RPF-VECTOR] Wijnands, I., Boers, A., and E. Rosen, "The Reverse Path Forwarding (RPF) Vector TLV", RFC 5496, 2009.

### 7.2. Informative References

- [IGMP-SNOOP] Christensen, M., Kimball, K., and F. Solensky, "Considerations for IGMP and MLD Snooping PEs", RFC 4541, 2006.

## [VPLS-BGP]

Kompella, K. and Y. Rekhter, "Virtual Private LAN Service using BGP for Auto-Discovery and Signaling", RFC 4761, 2007.

## [VPLS-LDP]

Lasserre, M. and V. Kompella, "Virtual Private LAN Services using LDP Signaling", RFC 4762, 2007.

## [VPLS-MCAST-REQ]

Kamite, Y., Wada, Y., Serbest, Y., Morin, T., and L. Fang, "Requirements for Multicast Support in Virtual Private LAN Services", RFC 5501, 2009.

## [VPLS-MCAST-TREES]

Aggarwal, R., Kamite, Y., Fang, L., and Y. Rekhter, "Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-11, Work in Progress.

## Appendix A. PIM-BIDIR Thoughts

This section describes some guidelines that may be used to preserve PIM-BIDIR functionality in combination with Pim Snooping.

In order to preserve PIM-BIDIR Pim snooping routers need to set up forwarding states so that :

- o on the RPL all traffic is forwarded to all Rport(N)
- o on any other interface traffic is always forwarded to the DF

The information needed to setup these states may be obtained by :

- o determining the mapping between group(range) and RP
- o snooping and storing DF election information
- o determining where the RPL is, this could be achieved by static configuration, or by combining the information mentioned in previous bullets.

### A.1. PIM-BIDIR Data Forwarding Rules

The PIM-BIDIR Snooping forwarding rules are defined below in pseudocode:

```
BEGIN
  iif is the incoming port of the multicast packet.
  G is the Destination IP Address of the multicast packet.

  If there is forwarding state for G
  Then
    OutgoingPortList = olist(G)
  Else
    OutgoingPortList = UserDefinedPortList
  Endif

  If iif is an AC
  Then
    OutgoingPortList = OutgoingPortList (-) iif
  Else
    ## iif is a PW
    OutgoingPortList = OutgoingPortList (-) PWPorts
  Endif

  Forward the packet to OutgoingPortList.
END

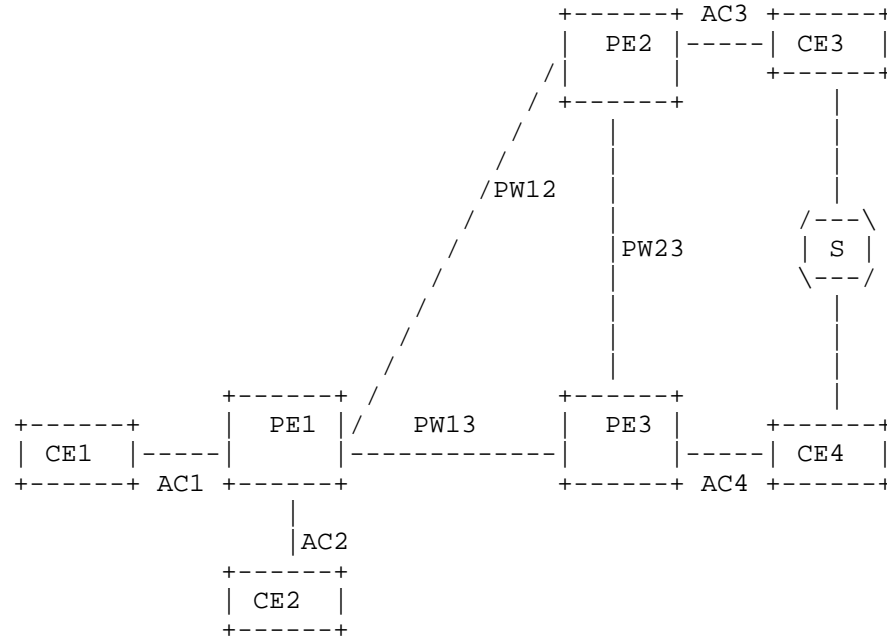
If there is forwarding state for G, then forward the packet to
olist(G) while observing the general rules above in Section 2.12

[PIM-BIDIR] specifies how olist(G) is constructed.
```

## Appendix B. Example Network Scenario

Let us consider the scenario in Figure 3.

## An Example Network for Triggering Assert



In the examples below,  $JT(\text{Port}, S, G, N)$  is the downstream Join Expiry Timer on the specified Port for the  $(S, G)$  with upstream neighbor  $N$ .

## B.1. Pim Snooping Example

In the network depicted in Figure 3,  $S$  is the source of a multicast stream  $(S, G)$ .  $CE1$  and  $CE2$  both have two ECMP routes to reach the source.

1.  $CE1$  Sends a  $Join(S, G)$  with  $UpstreamNeighbor(S, G) = CE3$ .
2.  $PE1$  snoops on the  $Join(S, G)$  and builds forwarding states since it is received on an AC. It also floods the  $Join(S, G)$  in the VPLS.  $PE2$  snoops on the  $Join(S, G)$  and builds forwarding state since the  $Join(S, G)$  is targeting a neighbor residing on an AC.  $PE3$  does not create forwarding state for  $(S, G)$  because this is a PW-only join and there is neither existing  $(*, G)$  state with an AC in  $UpstreamPorts(*, G)$  nor an existing  $(S, G)$  state with an AC in  $UpstreamPorts(S, G)$ . Both  $PE2$  and  $PE3$  will also flood the  $Join(S, G)$  in the VPLS

The resulting states at the PEs is as follows:

At  $PE1$ :

```

JT(AC1, S, G, CE3)          = JP_HoldTime
UpstreamNeighbors(S, G)    = { CE3 }

```

```

UpstreamPorts(S,G)      = { PW12 }
OutgoingPortList(S,G)   = { AC1, PW12 }

```

At PE2:

```

JT(PW12,S,G,CE3)       = JP_HoldTime
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)      = { AC3 }
OutgoingPortList(S,G)   = { PW12, AC3 }

```

At PE3:

No (S,G) state

3. The multicast stream (S,G) flows along CE3 -> PE2 -> PE1 -> CE1
4. Now CE2 sends a Join(S,G) with Upstream Neighbor(S,G) = CE4.
5. All PEs snoop on the Join(S,G), build forwarding state and flood the Join(S,G) in the VPLS. Note that for PE2 even though this is a PW-only join, forwarding state is built on this Join(S,G) since PE2 has existing (S,G) state with an AC in UpstreamPorts(S,G)

The resulting states at the PEs:

At PE1:

```

JT(AC1,S,G,CE3)        = active
JT(AC2,S,G,CE4)        = JP_HoldTime
UpstreamNeighbors(S,G) = { CE3, CE4 }
UpstreamPorts(S,G)      = { PW12, PW13 }
OutgoingPortList(S,G)   = { AC1, PW12, AC2, PW13 }

```

At PE2:

```

JT(PW12,S,G,CE4)       = JP_HoldTime
JT(PW12,S,G,CE3)       = active
UpstreamNeighbors(S,G) = { CE3, CE4 }
UpstreamPorts(S,G)      = { AC3, PW23 }
OutgoingPortList(S,G)   = { PW12, AC3, PW23 }

```

At PE3:

```

JT(PW13,S,G,CE4)       = JP_HoldTime
UpstreamNeighbors(S,G) = { CE4 }
UpstreamPorts(S,G)      = { AC4 }
OutgoingPortList(S,G)   = { PW13, AC4 }

```

6. The multicast stream (S,G) flows into the VPLS from the two CEs CE3 and CE4. PE2 forwards the stream received from CE3 to PW23 and PE3 forwards the stream to AC4. This facilitates the CE routers to trigger assert election. Let us say CE3 becomes the assert winner.
7. CE3 sends an Assert message to the VPLS. The PEs flood the Assert message without examining it.

8. CE4 stops sending the multicast stream to the VPLS.
9. CE2 notices an RPF change due to Assert and sends a Prune(S,G) with Upstream Neighbor = CE4. CE2 also sends a Join(S,G) with Upstream Neighbor = CE3.
10. All the PEs start a prune-pend timer on the ports on which they received the Prune(S,G). When the prune-pend timer expires, all PEs will remove the downstream (S,G,CE4) states.

Resulting states at the PEs:

At PE1:

JT(AC1,S,G,CE3)	= active
UpstreamNeighbors(S,G)	= { CE3 }
UpstreamPorts(S,G)	= { PW12 }
OutgoingPortList(S,G)	= { AC1, AC2, PW12 }

At PE2:

JT(PW12,S,G,CE3)	= active
UpstreamNeighbors(S,G)	= { CE3 }
UpstreamPorts(S,G)	= { AC3 }
OutgoingPortList(S,G)	= { PW12, AC3 }

At PE3:

JT(PW13,S,G,CE3)	= JP_HoldTime
UpstreamNeighbors(S,G)	= { CE3 }
UpstreamPorts(S,G)	= { PW23 }
OutgoingPortList(S,G)	= { PW13, PW23 }

Note that at this point at PE3, since there is no AC in OutgoingPortList(S,G) and no (\*,G) or (S,G) state with an AC in UpstreamPorts(\*,G) or UpstreamPorts(S,G) respectively, the existing (S,G) state at PE3 can also be removed. So finally:

At PE3:

No (S,G) state

Note that at the end of the assert election, there should be no duplicate traffic forwarded downstream and traffic should flow only on the desired path. Also note that there are no unnecessary (S,G) states on PE3 after the assert election.

## B.2. PIM Proxy Example with (S,G) / (\*,G) interaction

In the same network, let us assume CE4 is the Upstream Neighbor towards the RP for G.

JPST(S,G,N) is the JP sending timer for the (S,G) with upstream neighbor N.



1. CE1 Sends a Join(S,G) with Upstream Neighbor(S,G) = CE3.
2. PE1 consumes the Join(S,G) and builds forwarding state since the Join(S,G) is received on an AC.

PE2 consumes the Join(S,G) and builds forwarding state since the Join(S,G) is targeting a neighbor residing on an AC.

PE3 consumes the Join(S,G) but does not create forwarding state for (S,G) since this is a PW-only join and there is neither existing (\*,G) state with an AC in UpstreamPorts(\*,G) nor an existing (S,G) state with an AC in UpstreamPorts(S,G)

The resulting states at the PEs is as follows:

PE1 states:

```
JT(AC1,S,G,CE3)      = JP_HoldTime
JPST(S,G,CE3)        = t_periodic
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)    = { PW12 }
OutgoingPortList(S,G) = { AC1, PW12 }
```

PE2 states:

```
JT(PW12,S,G,CE3)     = JP_HoldTime
JPST(S,G,CE3)        = t_periodic
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)    = { AC3 }
OutgoingPortList(S,G) = { PW12, AC3 }
```

PE3 states:

No (S,G) state

Joins are triggered as follows:

PE1 triggers a Join(S,G) targeting CE3. Since the Join(S,G) was received on an AC and is targeting a neighbor that is residing across a PW, the triggered Join(S,G) is sent on all PWs.

PE2 triggers a Join(S,G) targeting CE3. Since the Joins(S,G) is targeting a neighbor residing on an AC, it only sends the join on AC3.

PE3 ignores the Join(S,G) since this is a PW-only join and there is neither existing (\*,G) state with an AC in UpstreamPorts(\*,G) nor an existing (S,G) state with an AC in UpstreamPorts(S,G)

3. The multicast stream (S,G) flows along CE3 -> PE2 -> PE1 -> CE1.
4. Now let us say CE2 sends a Join(\*,G) with UpstreamNeighbor(\*,G) = CE4.

5. PE1 consumes the Join(\*,G) and builds forwarding state since the Join(\*,G) is received on an AC.

PE2 consumes the Join(\*,G) and though this is a PW-only join, forwarding state is build on this Join(\*,G) since PE2 has existing (S,G) state with an AC in UpstreamPorts(S,G). However, since this is a PW-only join, PE2 only adds the PW towards PE3 (PW23) into UpstreamPorts(\*,G) and hence into OutgoingPortList(\*,G). It does not add the PW towards PE1 (PW12) into OutgoingPortsList(\*,G)

PE3 consumes the Join(\*,G) and builds forwarding state since the Join(\*,G) is targeting a neighbor residing on an AC.

The resulting states at the PEs is as follows:

PE1 states:

```
JT(AC1,* ,G,CE4)      = JP_HoldTime
JPST(* ,G,CE4)        = t_periodic
UpstreamNeighbors(* ,G) = { CE4 }
UpstreamPorts(* ,G)    = { PW13 }
OutgoingPortList(* ,G) = { AC2, PW13 }

JT(AC1,S,G,CE3)       = active
JPST(S,G,CE3)         = active
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)    = { PW12 }
OutgoingPortList(S,G) = { AC1, PW12, PW13 }
```

PE2 states:

```
JT(PW12,* ,G,CE4)     = JP_HoldTime
UpstreamNeighbors(* ,G) = { CE4 }
UpstreamPorts(G)       = { PW23 }
OutgoingPortList(* ,G) = { PW23 }

JT(PW12,S,G,CE3)      = active
JPST(S,G,CE3)         = active
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)    = { AC3 }
OutgoingPortList(S,G) = { PW12, AC3, PW23 }
```

PE3 states:

```
JT(PW13,* ,G,CE4)     = JP_HoldTime
JPST(* ,G,CE4)        = t_periodic
UpstreamNeighbors(* ,G) = { CE4 }
UpstreamPorts(* ,G)    = { AC4 }
OutgoingPortList(* ,G) = { PW13, AC4 }
```

Joins are triggered as follows:

PE1 triggers a Join(\*,G) targeting CE4. Since the Join(\*,G) was received on an AC and is targeting a neighbor that is residing across a PW, the triggered Join(S,G) is sent on all PWs.

PE2 does not trigger a Join(\*,G) based on this join since this is a PW-only join.

PE3 triggers a Join(\*,G) targeting CE4. Since the Join(\*,G) is targeting a neighbor residing on an AC, it only sends the join on AC4.

6. In case traffic is not flowing yet (i.e. step 3 is delayed to come after step 6) and in the interim JPST(S,G,CE3) on PE1 expires, causing it to send a refresh Join(S,G) targeting CE3, since the refresh Join(S,G) is targeting a neighbor that is residing across a PW, the refresh Join(S,G) is sent on all PWs.

7. Note that PE1 refreshes its JT timer based on reception of refresh joins from CE1 and CE2

PE2 consumes the Join(S,G) and refreshes the JT(PW12,S,G,CE3) timer.

PE3 consumes the Join(S,G). It also builds forwarding state on this Join(S,G), even though this is a PW-only join, since now PE2 has existing (\*,G) state with an AC in UpstreamPorts(\*,G). However, since this is a PW-only join, PE3 only adds the PW towards PE2 (PW23) into UpstreamPorts(S,G) and hence into OutgoingPortList(S,G). It does not add the PW towards PE1 (PW13) into OutgoingPortList(S,G).

PE3 States:

```

JT(PW13,*,G,CE4)           = active
JPST(S,G,CE4)              = active
UpstreamNeighbors(*,G)     = { CE4 }
UpstreamPorts(*,G)         = { AC4 }
OutgoingPortList(*,G)      = { PW13, AC4 }

JT(PW13,S,G,CE3)           = JP_HoldTime
UpstreamNeighbors(*,G)     = { CE3 }
UpstreamPorts(*,G)         = { PW23 }
OutgoingPortList(*,G)      = { PW13, AC4, PW23 }

```

Joins are triggered as follows:

PE2 already has (S,G) state, so it does not trigger a Join(S,G) based on reception of this refresh join.

PE3 does not trigger a Join(S,G) based on this join since this is a PW-only join.

8. The multicast stream (S,G) flows into the VPLS from the two

CEs, CE3 and CE4. PE2 forwards the stream received from CE3 to PW12 and PW23. At the same time PE3 forwards the stream received from CE4 to PW13 and PW23.

The stream received over PW12 and PW13 is forwarded by PE1 to AC1 and AC2.

The stream received by PE3 over PW23 is forwarded to AC4. The stream received by PE2 over PW23 is forwarded to AC3. Either of these facilitates the CE routers to trigger assert election.

9. CE3 and/or CE4 send(s) Assert message(s) to the VPLS. The PEs flood the Assert message(s) without examining it.
10. CE3 becomes the (S,G) assert winner and CE4 stops sending the multicast stream to the VPLS.
11. CE2 notices an RPF change due to Assert and sends a Prune(S,G,rpt) with Upstream Neighbor = CE4.
12. PE1 consumes the Prune(S,G,rpt) and since PruneDesired(S,G,Rpt,CE4) is TRUE, it triggers a Prune(S,G,rpt) to CE4. Since the prune is targeting a neighbor across a PW, it is sent on all PWs.

PE2 consumes the Prune(S,G,rpt) and does not trigger any prune based on this Prune(S,G,rpt) since this was a PW-only prune.

PE3 consumes the Prune(S,G,rpt) and since PruneDesired(S,G,rpt,CE4) is TRUE it sends the Prune(S,G,rpt) on AC4.

PE1 states:

```

JT(AC2,*,G,CE4)           = active
JPST(*,G,CE4)             = active
UpstreamNeighbors(*,G)    = { CE4 }
UpstreamPorts(*,G)        = { PW13 }
OutgoingPortList(*,G)     = { AC2, PW13 }

JT(AC2,S,G,CE4)           = JP_Holdtime with FLAG sgrpt prune
JPST(S,G,CE4)             = none, since this is sent along
                           with the Join(*,G) to CE4 based
                           on JPST(*,G,CE4) expiry
UpstreamPorts(S,G,rpt)    = { PW13 }
UpstreamNeighbors(S,G,rpt) = { CE4 }

JT(AC1,S,G,CE3)           = active
JPST(S,G,CE3)             = active
UpstreamNeighbors(S,G)    = { CE3 }
UpstreamPorts(S,G)        = { PW12 }

```

OutgoingPortList(S,G) = { AC1, PW12, AC2 }

At PE2:

```
JT(PW12,*,G,CE4)      = active
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(*,G)    = { PW23 }
OutgoingPortList(*,G) = { PW23 }

JT(PW12,S,G,CE4)      = JP_Holdtime with FLAG sgrpt prune
JPST(S,G,CE4)          = none, since this was created
                        off a PW-only prune
UpstreamPorts(S,G,rpt) = { PW23 }
UpstreamNeighbors(S,G,rpt) = { CE4 }

JT(PW12,S,G,CE3)      = active
JPST(S,G,CE3)          = active
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)     = { AC3 }
OutgoingPortList(*,G)  = { PW12, AC3 }
```

At PE3:

```
JT(PW13,*,G,CE4)      = active
JPST(*,G,CE4)          = active
UpstreamNeighbors(*,G) = { CE4 }
UpstreamPorts(*,G)     = { AC4 }
OutgoingPortList(*,G)  = { PW13, AC4 }

JT(PW13,S,G,CE4)      = JP_Holdtime with S,G,rpt prune flag
JPST(S,G,CE4)          = none, since this is sent along
                        with the Join(*,G) to CE4 based
                        on JPST(*,G,CE4) expiry
UpstreamNeighbors(S,G,rpt) = { CE4 }
UpstreamPorts(S,G,rpt)  = { AC4 }

JT(PW13,S,G,CE3)      = active
JPST(S,G,CE3)          = none, since this state is
                        created by PW-only join
UpstreamNeighbors(S,G) = { CE3 }
UpstreamPorts(S,G)     = { PW23 }
OutgoingPortList(S,G)  = { PW23 }
```

Even in this example, at the end of the (S,G) / (\*,G) assert election, there should be no duplicate traffic forwarded downstream and traffic should flow only to the desired CEs.

However, the reason we don't have duplicate traffic is because one of the CEs stops sending traffic due to assert, not because we don't

have any forwarding state in the PEs to do this forwarding.

Authors' Addresses

Olivier Dornon  
Alcatel-Lucent  
50 Copernicuslaan  
Antwerp, B2018

Email: [olivier.dornon@alcatel-lucent.com](mailto:olivier.dornon@alcatel-lucent.com)

Jayant Kotalwar  
Alcatel-Lucent  
701 East Middlefield Rd.  
Mountain View, CA 94043

Email: [jayant.kotalwar@alcatel-lucent.com](mailto:jayant.kotalwar@alcatel-lucent.com)

Venu Hemige

Email: [vhemige@gmail.com](mailto:vhemige@gmail.com)

Ray Qiu  
Juniper Networks, Inc.  
1194 North Mathilda Avenue  
Sunnyvale, CA 94089

Email: [rqiujuniper.net](mailto:rqiujuniper.net)

Jeffrey Zhang  
Juniper Networks, Inc.  
10 Technology Park Drive  
Westford, MA 01886

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



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:

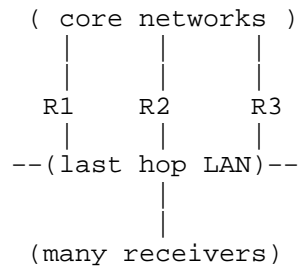


Figure 1: Last Hop Network

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.

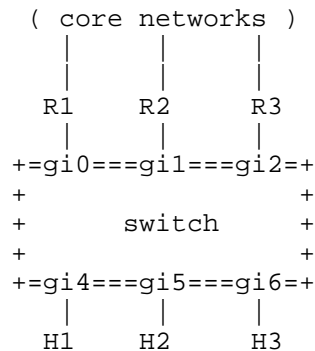


Figure 2: Last Hop Network with Ethernet Switch

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

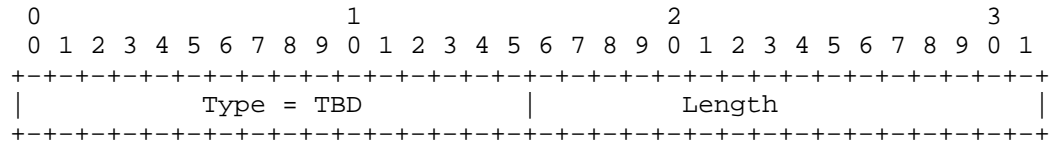


Figure 3: Capability Hello TLV

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

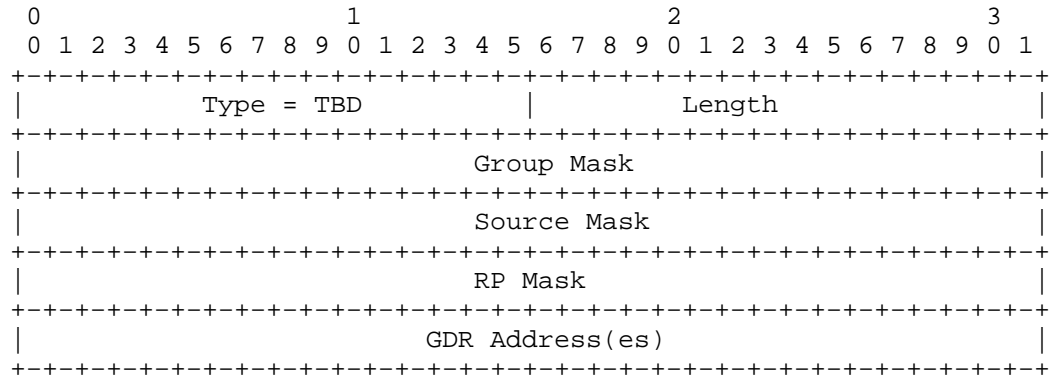


Figure 4: GDR Hello TLV



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



Network Working Group  
Internet-Draft  
Updates: 5384 (if approved)  
Intended status: Standards Track  
Expires: April 21, 2014

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

Hierarchical Join/Prune Attributes  
draft-ietf-pim-hierarchicaljoinattr-01.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 April 21, 2014.

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 . . . . .	2
2. Requirements Notation . . . . .	3
3. Hierarchical Join/Prune Attribute Definition . . . . .	3
4. Do Not Inherit (DNH) Join/Prune Attribute . . . . .	5
5. PIM Address Encoding Types . . . . .	6
6. Hierarchical Join/Prune Attribute Hello Option . . . . .	6
7. Security Considerations . . . . .	6
8. IANA Considerations . . . . .	7
9. Acknowledgments . . . . .	7
10. Normative References . . . . .	7
Authors' Addresses . . . . .	7

## 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 defines a new Join/Prune attribute to further control the scope of hierarchical attributes, as well as 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 5.

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. Do Not Inherit (DNH) Join/Prune Attribute

When considering the attributes for a specific source we inherit attributes from higher up in the hierarchy. As described above, if an attribute is specified at multiple levels, the one at the most specific level overrides the other instances of the attribute. However, we also want to be able to ignore attributes rather than overriding them. For instance if a given attribute and value should be specified for all sources but one, the attribute can be encoded in the Upstream Neighbor Address, but we need a way to specify that for the one source the attribute should not be applied. We do this by defining a new attribute called Do Not Inherit (DNH) that lists attribute types that should not be inherited from the less specific levels of the hierarchy. For the one source we add the DNH attribute specifying that the one attribute type should be ignored if present at the message wide or at the group level. Similarly the DNH attribute can be used at the group level to ignore certain attributes specified at the message wide level. If an attribute is ignored at a level, it can again be specified where needed at more specific levels.

##### Do Not Inherit 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|S| Attr Type | Length           | Ignore Type 1 | Ignore Type 2 |...
+-----+-----+-----+-----+-----+-----+-----+-----+...

```

F-bit (transitive) MUST be set to 0. This attribute MUST not be transitive. A router receiving a message using hierarchical attributes would decide whether this attribute is needed for efficient encoding and where when formatting a Join/Prune to send upstream, and this is independent of how a received Join/Prune was formatted.

E-bit (end of attributes) is set according to whether it is the last encoded attribute, see xref target="RFC5384"/>.

Attr Type = TBD1 is the type identifying this attribute.

Length denotes the length of the value field in octets.

The value field of the attribute consists of a list of attributes to not inherit. Each octet specifies one attribute type.

## 5. 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.

## 6. 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 = TBD2      |      OptionLength = 0      |
|                               |                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

OptionType = TBD2, 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.

## 7. Security Considerations

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

## 8. 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.

An assignment is needed from the Join attribute registry for the Do Not Inherit attribute. The string TBD1 needs to be replaced with the assigned value.

A new PIM Hello Option type needs to be assigned. The string TBD2 needs to be replaced with the permanently assigned value.

## 9. Acknowledgments

The authors would like to thank Siva Kollipara for providing feedback on the document.

## 10. 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

Network Working Group  
Internet Draft  
Category: Standard Track

L. Yong  
W. Hao  
D. Eastlake  
Huawei  
A. Qu  
J. Hudson  
Brocade

Expires: April 2014

October 18, 2013

ISIS Protocol Extension For Building Distribution Trees  
draft-yong-isis-ext-4-distribution-tree-01

Abstract

This document proposes an IS-IS protocol extension for automatically building bi-directional distribution trees to transport multi-destination traffic in an IP network.

Status of this document

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 April 18, 2014.

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.

## Table of Contents

1. Introduction.....	3
1.1. Conventions used in this document.....	4
2. IS-IS Protocol Extension.....	5
2.1. RTADDR sub-TLV.....	5
2.2. RTADDRV6 sub-TLV.....	6
2.3. The Group Address Sub-TLV.....	7
3. Procedures.....	8
3.1. Distribution Tree Computation.....	8
3.2. Parent Selection.....	8
3.3. Parallel Local Link Selection.....	9
3.4. Tree Selection for a Group.....	10
3.5. Pruning a Distribution Tree for a Group.....	10
3.6. RPF Mechanism.....	10
3.7. Forwarding Using a Pruned Distribution Tree.....	10
3.8. Local Forwarding at Edge Router.....	11
3.9. Distribution Tree across different IGP Levels.....	12
4. Backward Compatibility.....	13
5. Security Considerations.....	14
6. IANA Considerations.....	14
7. Acknowledgements.....	14
8. References.....	14
8.1. Normative References.....	14
8.2. Informative References.....	15



## 1. Introduction

Computer virtualization and cloud applications motivate the DC network virtualization technology [NVO3FRWK]. This technology decouples the end-points networking from the DC physical infrastructure network in terms of address space and configuration [NVO3FRWK].

DC network virtualization solutions are required to carry all types of traffic in today's DC physical networks including multi-destination traffic. It is also desirable to use an IP network as the DC underlying network for the overlay virtual networks [NVO3FRWK].

IP network technology does not yet support multi-destination traffic forwarding. A variety of Protocol Independent Multicast (PIM) solutions [RFC4601] [RFC5015] are designed to carry IP multicast traffic over IP networks. However the PIM solutions use their own hello protocol and hop-to-hop Join/Leave message so each router does not have global information about the receivers; in the PIM solution, the data packets could be forwarded unnecessarily to the Rendezvous Point (RP), and then get dropped there when no receiver at all or the sender and receivers for a multicast group are on the same branch towards the RP. This can unnecessarily consume network resources. Furthermore PIM solutions maintain a lot of soft-state, have intensive CPU utilization, and have additional convergence time, besides the IGP's, under a failure condition.

Although the PIM protocol is mature and has been deployed in IP networks, applying PIM to an IP network that supports the Network Virtualization can be an extremely challenging [MCASTISS]. For example, VXLAN [VXLAN] solutions require multicast support in the underlying network to simulate overlay L2 broadcast capability, where every edge node in an overlay virtual network (VN) is a multicast source and receiver. An overlay VN topology may be sparse and dynamic compared to the underlying IP network topology. Also a large number of overlay VNs may exist in a DC, which PIM solutions can't scale to.

Furthermore IP Overlay based network virtualization technology has been adopted by network vendors to create an automatically formed, self-healing, multi-service fabric to achieve the goal of a SDN capable fabric which is open, programmable, and elastic. Within the fabric, it is a closed IP network carrying all types of traffic, hence having one control plane protocol to support both uni-destination and multi-destination forwarding is highly desirable.

This document uses extensions to the IS-IS protocol to build a distribution tree for multi-destination traffic transport in an IP network. A router uses a Router Capability TLV to announce the tree root address and the multicast groups associated to the tree. With this information, routers in the IGP can compute rooted distribution trees by using the link state information, i.e. LSDB, and shortest path algorithm. Edge routers include information in their LSPs to announce their multicast group-memberships. Routers perform distribution tree pruning for each multicast group based on other router's group membership announcements. A router forwards the multi-destination traffic along the pruned tree.

In case that edge router needs to get the host membership of a multicast group, edge routers may use IGMP query messages [RFC3376] to inform the attached hosts and the hosts use IGMP report message to response with their interested multicast group(s).

In cases where the solution described in this document applies to the underlying network that transports overlay virtual networks [NVO3FRWK], mapping between an overlay multicast group and a underlying multicast group is necessary. Edge routers further need to perform packet encapsulation/decapsulation.

The benefits of this solution are 1) protocol convergence: use single protocol for both unicast and multicast traffic transport and get the same convergence time for unicast and multicast traffic. 2) multi-destination transport simplification: rely on the LSDB for computing a distribution tree and not run PIM hello protocol. 3) forwarding efficiency: no need to always forward the traffic to the RP; 4) better scalability: no need to maintain heavy PIM soft states. TRILL [RFC6325] has used IS-IS for both single destination and multi-destination packet transport, which proves the protocol capability of doing both.

#### 1.1. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

## 2. IS-IS Protocol Extension

## 2.1. RTADDR sub-TLV

This is a sub-TLV of the Router Capability TLV. Each RTADDR sub-TLV contains a root IPv4 address and multicast group addresses that associate to the tree. A router may use multiple RTADDR sub-TLVs to announce multiple root addresses and associated multicast groups with each root. RTADDR sub-TLV format is below.

```

+-----+
|Type=RTADDR| (1 byte)
+-----+
| Length | (1 byte)
+-----+
| Root IPv4 Address |
+-----+
|S| RESV| Topology ID | (2 byte)
+-----+
| Tree Priority | (1 byte)
+-----+
|Num of Groups | (1 byte)
+-----+
| Group Address (1) |
+-----+
| Group Mask (1) |
+-----+
~
+-----+
| GROUP Address (N) |
+-----+
| Group Mask (N) |
+-----+

```

Where:

Type: sub-TLV of Router Capability for RTADDR (TBD)

Length: variable depending on the number of associated groups

Root IPv4 Address: IPv4 Address for a root

S bit: If set, the rooted tree for single area only. Otherwise, the rooted tree crosses multiple areas.

RESV: 3 reserved bits. MUST be sent as zero and ignored on receipt.

Topology ID: This field carries a topology ID [RFC5120] or zero if topologies are not in use.

Tree Priority: An eight bit unsigned integer where larger magnitude means higher priority. Zero means no priority.

Num of Groups: the number of group addresses

Group Address: IPv4 Address for the group

Group Mask: multicast group range

One router may be the root for multiple trees. Each tree associates to a set of multicast groups. In this case, a router encodes multiple RTADDR sub-TLVs to announce root addresses, one for each root, in a router capability TLV. The group address/mask in different sub-TLVs can overlap. See section 3 for detail.

## 2.2. RTADDRV6 sub-TLV

This sub-TLV is used in an IPv6 network. It has the same format and usage except that the addresses are in IPv6.

```

+---+---+---+---+---+
|Type = RTADDRV6|          (1 byte)
+---+---+---+---+---+
|   Length       |          (1 byte)
+---+---+---+---+---+
|
+
|
+
|          Root IPv6 Address
|
+
+
+---+---+---+---+---+
|S|RESV|   Topology ID   |   (2 byte)
+---+---+---+---+---+
| Tree Priority |          (1 byte)
+---+---+---+---+---+
|Num of Groups  |          (1 byte)
+---+---+---+---+---+
|
+
|
+
|          Group IPv6 Address (1)
|
+
+
+---+---+---+---+---+
|
+
|
+
|          MASK(1)
|
+
+
+---+---+---+---+---+
~
+---+---+---+---+---+

```

### 2.3. The Group Address Sub-TLV

The Group Address TLV and a set of Group Address sub-TLVs are defined in RFC6326-bis [RFC6326BIS]. The GIP-ADDR and GIPV6-ADDR sub-TLVs are used in this solution. An edge router uses the GIP-ADDR sub-TLV or GIPV6-ADDR to announce its interested multicast groups.

The GIP-ADDR sub-TLV applies to an IPv4 network and GIPV6-ADDR sub-TLV for IPv6 network.

When using a GIP-ADDR or GIPV6-ADDR sub-TLV, the field VLAN-ID MUST set to zero and be ignored. Other field usage remains the same as [RFC6326BIS]

### 3. Procedures

When an operator selects a router as a distribution tree root, he/she configures the tree root address and associated multicast groups with the router. A tree root address can be an interface address or router loopback address. After the configuration, the router will include a RTADDR sub-TLV, inside a router capability TLV, where the tree root address and multicast groups are specified. If multiple trees are configured on the router, multiple RTADDR sub-TLVs are added in one router capability TLV to specify individual tree roots. For IPv4 network, RTADDR sub-TLV is used. For IPv6, RTADDRV6 sub-TLV is used. Note that the rest of document specifies the processes for an IPv4 network only. The processes for an IPv6 network are the same.

Operators may associate one multicast group to more than one tree for the redundancy purposes and use the tree priority to specify the primary tree preference. Section 3.2 describes the primary tree selection.

#### 3.1. Distribution Tree Computation

Upon receiving RTADDR sub-TLVs, routers track the tree roots and associated multicast groups. When the LSDB stabilizes, routers calculate all rooted trees according to the LSDB and shortest path algorithm.

One multicast group may associate to multiple trees. It is important that all the routers choose the same tree for a multicast group. Section 3.2 and 3.3 describes the tiebreaking rule for primary tree selection for a multicast group and parent selection in case of equal-cost to potential children.

#### 3.2. Parent Selection

It is important, when building a distribution tree, that all routers choose the same links for the tree. Therefore, when there are equal costs from a potential child node to possible parent nodes, all routers need to use the same tiebreakers. It is also desirable to

allow splitting of traffic on as many links as possible in such situations. TRILL [RFC6325] achieves this by defining multiple rooted trees and using the tiebreakers to enable nodes in these trees to choose different parents. This draft uses the same tiebreakers as TRILL [RFC6325].

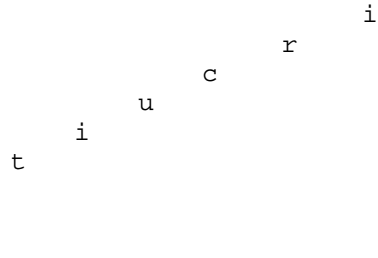
If there are  $k$  distribution trees in the network, when each router computes these trees, the  $k$  trees calculated are ordered and numbered from 0 to  $k-1$  in ascending order according to root IP addresses.

The tiebreaker rule is: When building the tree number  $j$ , remember all possible equal cost parents for router  $N$ . After calculating the entire "tree" (actually, directed graph), for each router  $N$ , if  $N$  has " $p$ " parents, then order the parents in ascending order according to the 7-octet IS-IS ID considered as an unsigned integer, and number them starting at zero. For tree  $j$ , choose  $N$ 's parent as choice  $j \bmod p$ .

### 3.3. Parallel Local Link Selection

If there are parallel point-to-point links between two routers, say  $R1$  and  $R2$ , these parallel links would be visible to  $R1$  and  $R2$ , but not to other routers. If this bundle of parallel links is included in a tree, it is important for  $R1$  and  $R2$  to decide which link to use; if the  $R1$ - $R2$  link is the branch for multiple trees, it is desirable to split traffic over as many link as possible. However the local link selection for a tree irrelevant to other Routers. Therefore, the tiebreaking algorithm need not be visible to any Routers other than  $R1$  and  $R2$ .

When there are  $L$  parallel links between  $R1$  and  $R2$  and they both are on  $K$  trees.  $L$  links are ordered from 0 to  $L-1$  in ascending order of  $C$



Circuit ID as associated with the adjacency by the router with the highest System ID, and  $K$  trees are ordered from 0 to  $K-1$  in ascending order of root IP addresses. The tiebreaker rule is: for tree  $k$ , select the link as choice  $k \bmod L$ .

Note that if multiple distribution trees are configured in a network or on a router, better load balance among parallel links through the tie-breaking algorithm can be achieved. Otherwise, if there is only one tree is configured, then only one link in parallel links can be used for the corresponding distribution tree. However, calculating and maintaining many trees is resource consuming. Operators need to balance between two.





### 3.4. Tree Selection for a Group

Routers receive one or more possible multicast group-range-to-tree mappings. Each mapping specifies a range of multicast groups. It is possible that a group-range is associated with multiple trees that may have the same or different priority. When a multicast group-range associates with more than one tree, all routers have to select the same tree for the group-range. The tiebreaker rules specified in PIM [RFC4601] are used. They are:

- o Perform longest match on group-range to get a list of trees.
- o Select the tree with highest priority.
- o If only one tree with the highest priority, select the tree for the group-range.
- o If multiple trees are with the highest priority, use the PIM hash function to choose one. PIM hash function is described in section 4.1.1 in RFC4601 [RFC4601].

### 3.5. Pruning a Distribution Tree for a Group

Routers prune the distribution tree for each associated multicast group, i.e. eliminating branches that have no potential downstream receivers. Multi-destination packets SHOULD only be forwarded on branches that are not pruned. The assumption here is that a multicast source is also a multicast receiver but a multicast receiver may not be a multicast source.

Routers prune the trees based on the groups specified in GRADD-TLV from edge routers. Routers maintain a list of adjacency interfaces that are on the pruned tree for a multicast group. Among these interfaces, one interface may be toward the tree-root router and other are toward the egress routers.

### 3.6. RPF Mechanism

For the further study.

### 3.7. Forwarding Using a Pruned Distribution Tree

Forwarding a multi-destination packet follows the pruned tree for the group that the packet belongs to. It is done as follows.

- o If the router receives a multi-destination packet with group IP address that does not associated with any tree, the packet MUST be dropped.
- o Else check if the link that the packet arrives on is one of the ports in the pruned distribution tree. If not, the packet MUST be dropped.
- o Else perform RPF checking (section 3.5). If it fails, the packet SHOULD be dropped.
- o Else the packet is forwarded onto all the adjacency interfaces in the list for the group except the interface where the packet receive.

### 3.8. Local Forwarding at Edge Router

Upon receiving a multi-destination packet, besides forwarding it along the pruned tree, an edge router may also need to forward the packet to the local hosts attached to it. This is referred to as local forwarding in this document.

The local group database is needed to keep track of the group membership of the router's directly attached network or host. Each entry in the local group database is a [group, network/host] pair, which indicates that the attached network has one or more hosts belonging to the multicast group. When receiving a multi-destination packet, the edge router forwards the packet to the network/host that match the [group, network/host] pair in the local group database.

The local group database is built through the operation of the IGMPv3 [RFC3376]. When an edge router becomes Designated Router on an attached network, say N1, it starts sending periodic IGMPv3 Host Membership Queries on the network. Hosts then respond with IGMPv3 Host Membership Reports, one for each multicast group to which they belong. Upon receiving a Host Membership Report for a multicast group A, the router updates its local group database by adding/refreshing the entry [Group A, N1]. If at a later time Reports for Group A cease to be heard on the network, the entry is then deleted from the local group database. The Designated further sends the LSP message with GRADDR sub-TLV to inform other routers about the group memberships in the local group database. A router MUST ignore Host Membership Reports received on those networks where the router has not been elected Designated Router.

When the solution described in this document applies to the underlying network that transports overlay virtual networks [NVO3FRWK], A Designated Router further necessarily maintains the mapping between an overlay multicast group and a underlying multicast group, and performs packet encapsulation/descapsulation upon receiving a packet from host or the underlying network. Mapping between an overlay multicast group and a underlying multicast group can be manually configured, automatically generated by an algorithm, or dynamically informed at a Designated Router. The same edge router should be selected as the Designated Router for the overlay multicast group and underlying multicast group that are associated. The mapping method is beyond the scope of this document.

### 3.9. Distribution Tree across different IGP Levels

An IGP (Interior Gateway Protocol) network may be designed as a multi-area network for the scalability, faster-convergence. Multicast sources and listeners may be in the same or different areas. The former is a special case of the latter. To support multi-destination transport over multi-areas, it is necessary to build a distribution tree across areas and prune the tree based on the listener locations, i.e. interested edge routers that may reside in different areas.

For an IS-IS multi-area network, there are level1 and level2 routers as well as level1/2 (border) routers. A level1 router only has the router/topology information in its area. A level2 router has router/topology information in level2 area as well as router information in level1 areas. A border router participates in both level1 and level2 areas and has the router/topology information in both attached areas but maintain two separated LSDBs. Traffic from one area to another area must traverse through a border router. It is possible to have more than one border router between two areas for resilience.

To build a distribution tree across mutli-areas, an operator can select a tree-root node for a set of multicast groups. The node can be in level1 area or level2 area. All the nodes including border nodes in the area compute the distribution tree as described in section 3.1-3.4. Border routers automatically select a designated forwarder for the multicast groups associated to the tree (see below). The border router selected as designate forwarder (DF) announces itself as the tree root in the adjacent area if the S bit in the RTADDR TLV is clear. The nodes in the adjacent area will compute the distribution tree in the same way. Note that a border router may be the tree-root in the adjacent area for the multicast groups that may associate with different trees. If S bit in the

RTADDR TLV is set, the rooted distribution tree is only built in the area where the root node resides.

The document specifies following additional rules for a border router that supports the multicast mechanism described in this document. The rules apply to the case of the distribution tree across multiple areas.

If a border router is selected as designated forwarder in adjacent area for a set of multicast groups, it should perform following:

- o It MUST track the group-memberships in its participated areas.
- o It MUST send a summary group membership of one area to the adjacent area as of an edge router.
- o It performs the pruning process in each area, respectively, based on the received group-membership LSPs from that area.
- o When receiving multicast traffic from one area, it forwards the packet along the pruned tree into the adjacent area.
- o Optional performs reverse path forwarding check

If a border router is not selected as the designated forwarder for the multicast groups, it MUST do following:

- o It should not propagate the group-membership information of one area to any other areas.
- o It should not forward multicast group traffic to another adjacent area.

The method of selecting a border router as the designated forwarder of multicast group traffic will be addressed in next version of this document.

#### 4. Backward Compatibility

If a router does not support the distribution tree function described in this document, distribution tree computation MUST NOT include this router. This may result the incomplete tree. An operator can build a tunnel between two routers, which allows a single rooted tree to be built. How to build the tunnel is outside scope of this document.

## 5. Security Considerations

Coming soon.

## 6. IANA Considerations

The document requires two new sub-TLVs, RTADDR and RTADDRV6 for the Router Capability TLV in IANA registry.

## 7. Acknowledgements

Authors like to thank Mike McBride and Linda Dunbar for their valuable inputs.

## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, March 1997.
- [RFC3376] Cain B., etc, ''Internet Group Management Protocol, Version 3'', rfc4604, October 2002
- [RFC4601] Fenner, B., et al, ''Protocol Independent multicast -  
Sparse Mode (PIM-SM): Protocol Specification'', rfc4601,  
August 2006
- [RFC5015] Handley, M., et al, ''Bidirectional Protocol Independent Multicast (BIDIR-PIM'', rfc5015, October 2007
- [RFC5120] Przygienda, T., et al, ''M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)'', rfc5120, February 2008
- [RFC6325] Perlman, R., et al, ''Routing Bridges (RBridges): Base Protocol Specification'', RFC6325, July 2011
- [RFC6326BIS] Eastlake, D., et al, ''Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS'', draft-ietf-isis-rfc6326bis-01, work in progress

## 8.2. Informative References

- [MCASTISS] Ghanvani, A., ''Multicast Issues in Networks Using NVO3'', draft-ghanwani-nvo3-mcast-issues-00, work in progress
- [NVO3FRWK] Lasserre, M., ''Framework for DC Network Virtualization'', draft-ietf-nvo3-framework-03.txt, work in progress.
- [VXLAN] Mahalingam, M., Dutt, D., etc, ''VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks'', draft-mahalingam-dutt-dcops-vxlan-05.txt, work in progress

### Authors' Addresses

Lucy Yong  
Huawei USA  
5340 Legacy Drive  
Plano, TX 75025 USA

Phone: 469-277-5837  
Email: lucy.yong@huawei.com

Weiguo Hao  
Huawei Technologies  
101 Software Avenue,  
Nanjing 210012  
China

Phone: +86-25-56623144  
Email: haoweiguo@huawei.com

Donald Eastlake  
Huawei  
155 Beaver Street  
Milford, MA 01757 USA

Phone: +1-508-333-2270  
EMail: d3e3e3@gmail.com

Andrew Qu  
Brocade  
130 Holger Way

San Jose, CA 95134 USA

Email: laodulaodu@gmail.com

Jon Hudson

Brocade

130 Holger Way

San Jose, CA 95134 USA

Phone: +1-408-333-4062

Email: jon.hudson@gmail.com





Protocol Independent Multicast  
Internet-Draft  
Updates: 5015 (if approved)  
Intended status: Standards Track  
Expires: April 19, 2014

Z. Zhang  
K. Windisch  
J. A. Gralak  
Juniper Networks, Inc.  
October 16, 2013

PIM-Bidir RPL Resiliency  
draft-zzhang-pim-bidir-rpl-resiliency-00.txt

## Abstract

With PIM-Bidir, the RPA does not have to be associated with a router. Rather, it only needs to be a routable address on a RPL (typically a multi-access network). Such a scenario is commonly referred as Phantom RPA. This achieves RP resiliency to some extent, because the "RP" will not fail. However, if the RPL itself partitions, traffic converged to one partition will not be able to reach other parts of the network where joins converge to the other partitions of the RPL.

This document proposes simple procedures, which does not require signaling extensions, to achieve RPL resiliency.

## 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 April 19, 2014.

## 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

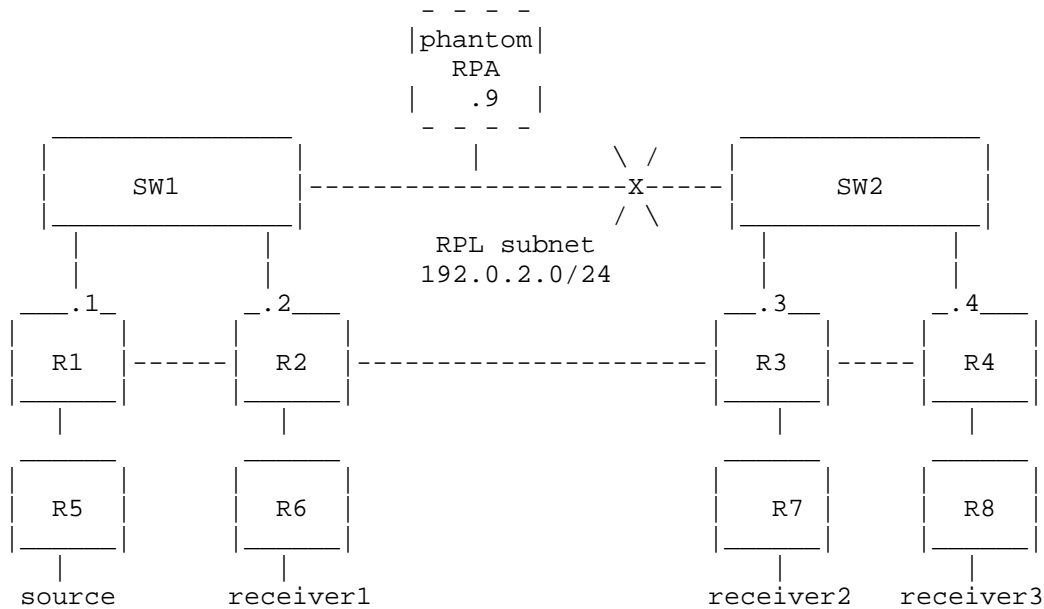
1. Introduction . . . . .	2
1.1. Problem Description . . . . .	2
1.2. Motivations . . . . .	3
1.3. Proposed Solutions . . . . .	4
2. Operations . . . . .	5
2.1. Modified PIM-Bidir Procedures . . . . .	5
2.2. Detect partitioning and elect active partition . . . . .	6
2.2.1. Using Host Routes advertised by any protocol . . . . .	6
2.2.2. Using Link State Routing protocol . . . . .	6
2.2.3. Comparison between the two detection and election methods . . . . .	8
3. IANA Considerations . . . . .	8
4. Security Considerations . . . . .	8
5. Contributors . . . . .	8
6. Acknowledgements . . . . .	9
7. References . . . . .	9
7.1. Normative References . . . . .	9
7.2. Informative References . . . . .	9

## 1. Introduction

### 1.1. Problem Description

The problem with partitioned RPL is that routers on the RPL still expect traffic to be exchanged over the RPL to reach other parts of the network, even though that won't happen across the RPL partitions.

This can be illustrated by Figure 1. The RPL is served by two interconnected switches and if the link between the switches breaks, R1~R4 will all continue to treat the link as RPL, and terminate the joins. R3~4 continue to expect traffic injected by R5 to arrive on the RPL link, instead of sending joins to R2.



RPL partition caused by the inter-switch link failure.

Figure 1

## 1.2. Motivations

The importance of ensuring traffic reachability in spite of RPL partitioning is obvious. Additionally, [I-D.wijnands-pim-source-discovery-bsr] provides a perfect example of PIM-Bidir as a solution once the partitioning problem is solved.

[I-D.wijnands-pim-source-discovery-bsr] proposes to extend BSR to flood source information so that routers connecting to receivers can send (s,g) SPT joins, bypassing the RTP->SPT switch. It points out that the solution is not suitable "for applications with strong dependency on the initial packet(s)" and PIM-Bidir [RFC5015] should be used for that. However, PIM-Bidir is not suitable where high resiliency is required, unless the partitioning problem is resolved.

[I-D.wijnands-pim-source-discovery-bsr] also raises a question whether BSR should be extended to a generic flooding mechanism for opaque information. Due to the way BSR flooding is done, while it is acceptable to flood group-to-rp mapping, it becomes inefficient to flood large amount of data. PIM-Bidir can be used as a generic protocol for efficient many-to-many data distribution and solving the partitioning problem enables the same level of resiliency as BSR flooding.

### 1.3. Proposed Solutions

This problem can be solved as follows:

- o Routers on the RPL detect RPL partitioning, elect an active partition to continue function as RPL, and stop treating the inactive partitions as RPL.
- o All routers route joins and traffic towards the active partition.

For the first task, this document specifies two methods to detect RPL partitioning and elect an active partition. For the second task, a host route to the RPA can be announced by the routers on the active partition.

This solution not only addresses RPL partitioning, it can also be used to mitigate the impact of network partitioning (where a part of network may be completely separated from the rest) by intentionally placing RPL segments into different parts of the network, as illustrated on Figure 2. This is called Anycast RPL in this document, because the segments will all have the same subnet. With that, only one segment will be active and treated as RPL before the network partitions.

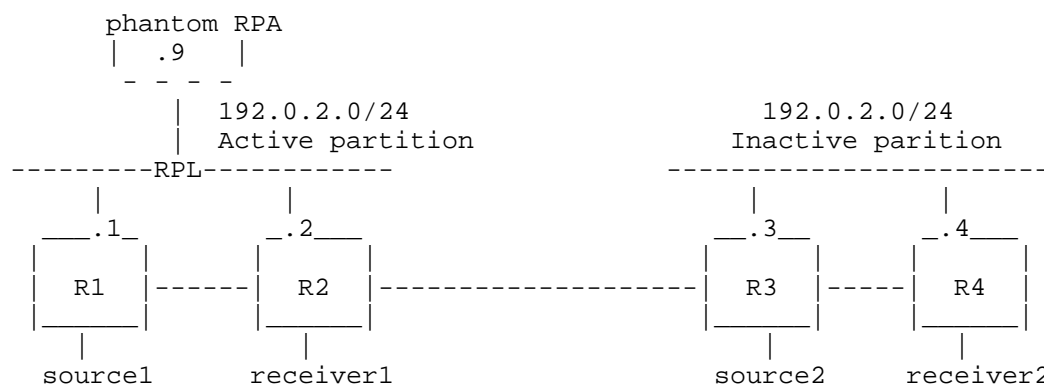


Figure 2

When the network separates into completely disjoint partitions, see figure Figure 3, each partition may have their own active RPL so intra-partition traffic will continue to flow. In the extreme case, all routers can be put onto RPL segments, making the network extremely resilient from PIM-Bidir point of view.

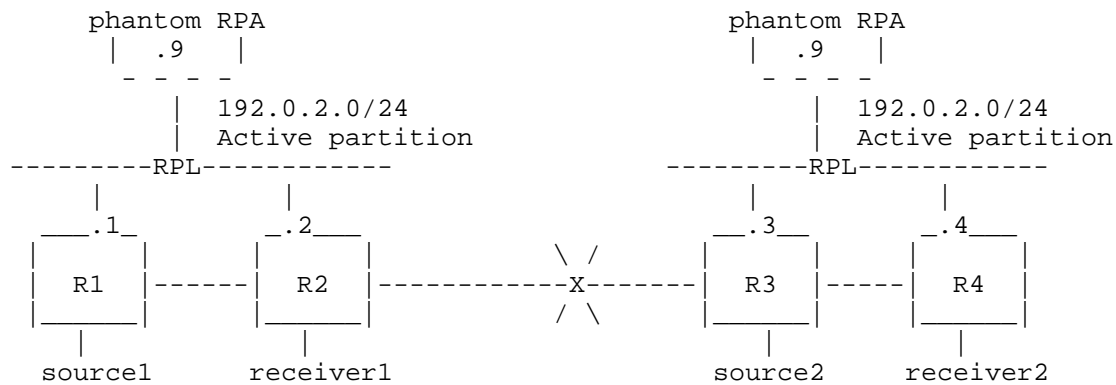


Figure 3

For simplicity and practicality, this document assumes that the RPA does not belong to any router on the RPL. Such a scenario is commonly referred as phantom RPA. These procedures MUST NOT be used when the RPA is an address belonging to a router.

## 2. Operations

### 2.1. Modified PIM-Bidir Procedures

A PIM router treats a link as RPL when the following two conditions are all met:

- o [existing] The route towards a RPA is directly over the link
- o [new] The router is in the elected active partition

Note that the active partition could be the one and only "partition" (when there is no RPL partitioning).

A router MUST advertise a host route to the RPA if and only if it treats a link as RPL. It MUST start the DF election on the link and treat it as a regular link when it stops treating a link as RPL. When it starts treating a link as RPL, it MUST stop the DF election.

The following sections specify two methods to detect partitioning and elect an active partition. Each elected active partition is

identified by one of the routers, and other routers determine if they are in the active partition by checking their neighborhood with the identifying router.

The neighborhood check can be done via either IGP mechanism (e.g. OSPF Hello) or PIM Hello (if used). In either case, fast neighborhood change detection SHOULD be used, e.g., via BFD or short Hello interval.

## 2.2. Detect partitioning and elect active partition

For the detection and election, each partition needs to be represented by one or more identifiers. This can be done by two methods.

### 2.2.1. Using Host Routes advertised by any protocol

In each partition, routers learn of each other by way of PIM Hellos. Of all the neighbors, the one with the lowest routable unicast interface address on the subnet MUST advertise a host route to the address itself, e.g. via a Stub Link in the OSPF Router LSA or a BGP NLRI. Optionally, to speed up convergence and facilitate make-before-break process, the one with the second lowest address or even all may do the same.

The host routes represent all partitions, potentially with N:1 mapping.

Routers on the RPL subnet find all the host routes that fall into the RPL subnet range, and select the one with the lowest address which itself not RPA address. That address identifies the active partition. Whenever such a host route is added or deleted, the election process is rerun.

### 2.2.2. Using Link State Routing protocol

When a Link State Routing protocol is used, the link states for the RPL subnet can be used. For example, with OSPF, each partition may have its own Network LSA for the same subnet, or in case of no Network LSA (there may be no DR or adjacency between the DR and a non-DR), each router on the partition will advertise a stub link in its Router LSA for the RPL subnet. Routers on the RPL subnet check all the reachable Network LSAs for the subnet and reachable Router LSAs that have a stub link for the subnet. The Network LSA with the lowest Advertising Router among all those Network LSAs, or in case of no Network LSAs the Router LSA with the lowest Advertising Router is selected to identify the active partition. If a Network LSA is selected, then a router is on the active partition if and only if it

originated the Network LSA, or it is a neighbor on the subnet with the Advertising Router. If a Router LSA is selected, then only the Advertising Router itself is on the active partition.

Whenever a corresponding Network LSA or stub link for the RPL subnet is added/deleted or its reachability changes, the election process is rerun.

The above procedure does not need any PIM/IGP signaling extensions, but only works if all the partitions are in the same area. That is sufficient to address RPL partitioning, but if it is desired to put Anycast RPLs in different areas, then IGP signaling extension is needed. Again using OSPF as an example:

- o When an Area Border Router (ABR) advertises a Type 3 Summary LSA into the backbone area B from a non-backbone area A for a RPL subnet that it learns in area A, the Summary LSA MUST carry, in a TLV according to [I-D.acee-ospfv3-lsa-extend](details TBD), the lowest Advertising Router of the reachable Network LSAs for the RPL Subnet, or in case of no Network LSAs, the lowest Advertising Router of reachable Router LSAs that have a stub link for the RPL subnet, plus the LSA Type. If the ABR belongs to multiple non-backbone areas and the RPL subnet is reachable in more than one of the areas, a single Summary LSA is originated. In that case, Advertising Router and LSA Type in the TLV is set according to the selected LSA in the area with the lowest Area ID.
- o When an ABR advertises a Type 3 Summary LSA into the non-backbone area A for a RPL subnet that it learns in the backbone area B, the Summary LSA MUST carry in a TLV according to [I-D.acee-ospfv3-lsa-extend] the Advertising Router of the LSA that identifies the elected active partition. If the LSA is a Network LSA or Router LSA, the LSA Type in the TLV is set accordingly. If it is a Summary LSA, the LSA Type is copied from the Summary LSA's TLV. The LSA Type is not really used, but included for consistency.
- o For the election process in the backbone area, Advertising Routers in the following ordered groups are compared, and the lowest Advertising Router in the first non-empty group is elected to identify the active partition.
  - \* Of reachable Network LSAs for the RPL subnet
  - \* Of reachable Router LSAs with stub link for the RPL subnet
  - \* Carried in the above mentioned TLV of reachable Summary LSAs for the RPL subnet, with Network LSA type in the TLV

- \* Carried in the above mentioned TLV of reachable Summary LSAs for the RPL subnet with Router LSA type in the TLV
- o In a non-backbone area, the following group order is used instead, so that the partitions in the backbone area are always preferred.
  - \* Carried in the above mentioned TLV of reachable Summary LSAs for the RPL subnet
  - \* Of reachable local Network LSAs for the RPLsubnet
  - \* Of reachable Router LSAs with stub link for the RPL subnet
- o To determine if a router is on the active partition, the router checks if the active partition is identified by a Summary LSA. If yes, the Advertising Router from the TLV is used. Otherwise, the Advertising Router of the identifying Network LSA or Router LSA is used. Then, the same neighborhood checking as in single area case is done to determine if the router is on the active partition.

#### 2.2.3. Comparison between the two detection and election methods

The Host Route method universally works with any routing protocol w/o any signaling changes, and can work across AS boundaries. However, it requires advertising additional host routes, and the election is purely based on address comparison.

The other method works only with Link State Routing protocols and only works intra-AS. It needs IGP signaling extensions if multiple RPL segments need to be intentionally placed in different areas. For OSPF, currently the extension is only considered for OSPFv3. On the other hand, it only needs a little additional signaling for the intentional inter-area Anycast RPL deployment, and the election prefers the RPL segments in the backbone area, which may be desired.

### 3. IANA Considerations

This document makes no request of IANA.

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

### 4. Security Considerations

This document does not introduce new security risks.

### 5. Contributors



## 6. Acknowledgements

## 7. References

### 7.1. Normative References

- [I-D.acee-ospfv3-lsa-extend]  
Lindem, A., Mirtorabi, S., Roy, A., and F. Baker, "OSPFv3 LSA Extendibility", draft-acee-ospfv3-lsa-extend-02 (work in progress), September 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.
- [RFC5015] Handley, M. ., Kouvelas, I. ., Speakman, T. ., and L. . Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

### 7.2. Informative References

- [I-D.wijnands-pim-source-discovery-bsr]  
Wijnands, I., Venaas, S., and M. Brig, "PIM flooding mechanism and source discovery", draft-wijnands-pim-source-discovery-bsr-03 (work in progress), July 2013.

## Authors' Addresses

Zhaohui (Jeffrey) Zhang  
Juniper Networks, Inc.  
10 Technology Park Drive  
Westford, MA 01886

EMail: zzhang@juniper.net

Kurt Windisch  
Juniper Networks, Inc.

EMail: kurtw@juniper.net

Jaroslav Adam Gralak  
Juniper Networks, Inc.

EMail: [jgralak@juniper.net](mailto:jgralak@juniper.net)