

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

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

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

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on August 29, 2013.

Copyright Notice

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

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

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

Table of Contents

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

1. Introduction

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

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

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

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

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

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

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [5].

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

Upstream interface (or selected upstream interface):

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

Downstream interface:

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

Candidate upstream interface:

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

Supported address prefix:

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

3. Per-Channel Load Balancing

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

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

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

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

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

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

4. Candidate Upstream Interface Configuration

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

4.1. Supported Address Prefix

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

(source address prefix, multicast address prefix)

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

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

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

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

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

4.2. Interface Priority

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

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

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

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

4.3. Default Interface

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

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

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

5. IANA Considerations

This document has no actions for IANA.

6. Security Considerations

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

7. Normative References

- [1] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [2] Liu, H., Cao, W., and H. Asaeda, "Lightweight Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Version 2 (MLDv2) Protocols", RFC 5790, February 2010.
- [3] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [4] Fenner, B., He, H., Haberman, B., and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.
- [5] Bradner, S., "Key words for use in RFCs to indicate requirement levels", RFC 2119, March 1997.

Authors' Addresses

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

Email: asaeda@nict.go.jp

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

Email: seiljeon@av.it.pt

PIM Working Group
Internet-Draft
Intended status: Experimental
Expires: January 4, 2014

LM. Contreras
Telefonica I+D
CJ. Bernardos
UC3M
JC. Zuniga
InterDigital
July 3, 2013

Extension of the MLD proxy functionality to support multiple upstream
interfaces
draft-contreras-pim-multiple-upstreams-00

Abstract

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

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

Status of this Memo

This Internet-Draft is submitted 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 January 4, 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	4
2. Terminology	4
3. Problem statement	4
4. Scenarios of applicability	7
4.1. Fixed network scenarios	7
4.1.1. Multicast wholesale offer for residential services	8
4.1.1.1. Requirements	8
4.1.2. Multicast resiliency	8
4.1.2.1. Requirements	8
4.1.3. Load balancing for multicast traffic in the metro segment	9
4.1.3.1. Requirements	9
4.1.4. Summary of the requirements needed for mobile network scenarios	9
4.2. Mobile network scenarios	10
4.2.1. Applicability to multicast listener mobility	10
4.2.1.1. Single MLD proxy instance on MAG	11
4.2.1.1.1. Requirements	11
4.2.1.2. Remote and local multicast subscription	11
4.2.1.2.1. Requirements	12
4.2.1.3. Dual subscription to multicast groups during handover	12
4.2.1.3.1. Requirements	13
4.2.2. Applicability to multicast source mobility	13
4.2.2.1. Support of remote and direct subscription in basic source mobility	13
4.2.2.1.1. Requirements	14
4.2.2.2. Direct communication between source and listener associated with distinct LMAs but on the same MAG	14
4.2.2.2.1. Requirements	15
4.2.2.3. Route optimization support in source mobility for remote subscribers	15
4.2.2.3.1. Requirements	15
4.2.3. Summary of the requirements needed for mobile network scenarios	16
5. Functional specification of an MLD proxy with multiple interfaces	18
6. Security Considerations	18
7. IANA Considerations	18
8. Acknowledgments	18
9. References	18
9.1. Normative References	18
9.2. Informative References	19
Appendix A. Basic support for multicast listener with PMIPv6	19
Authors' Addresses	21

1. Introduction

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

2. Terminology

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 [RFC2119].

This document uses the terminology defined in RFC4605 [RFC4605]. Specifically, the definition of Upstream and Downstream interfaces, which are reproduced here for completeness.

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

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

3. Problem statement

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

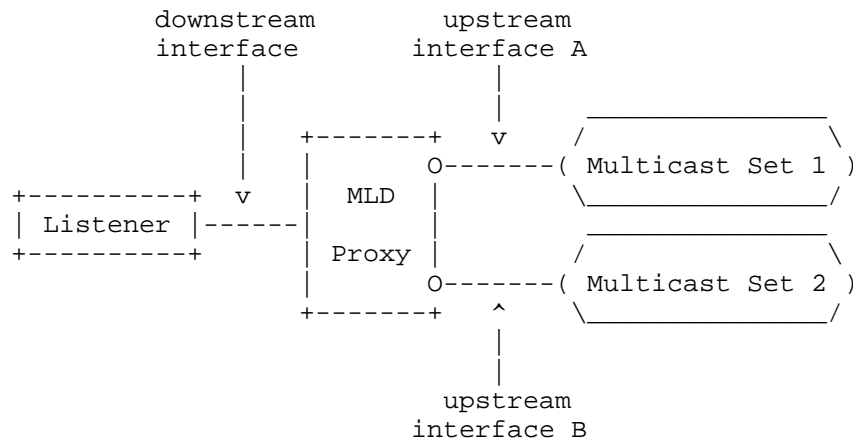


Figure 1: Concept of MLD proxy with multiple upstream interfaces

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

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

In the case of mobile networks, IP mobility services guarantee the continuity of the IP session while a Mobile Node (MN) changes its point of attachment. Proxy Mobile IPv6 (PMIPv6) RFC5213 [RFC5213] standardized a protocol that allows the network to manage the MN mobility without requiring specific support from the mobile terminal. The traffic to the MN is tunneled from the Home Network making use of two entities, one acting as mobility anchor, and the other as Mobility Access Gateway (MAG). Multicast support in PMIPv6 RFC6224 [RFC6224] implies the delivery of all the multicast traffic from the Home Network, via the mobility anchor. However, multicast routing optimization [I-D.ietf-multimob-pmipv6-ropt] could take advantage of an MLD proxy with multiple upstream interfaces by supporting the decision of subscribing a multicast content from the Home Network or from the local PMIPv6 domain if it is locally available. Figure 3

presents this scenario.

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

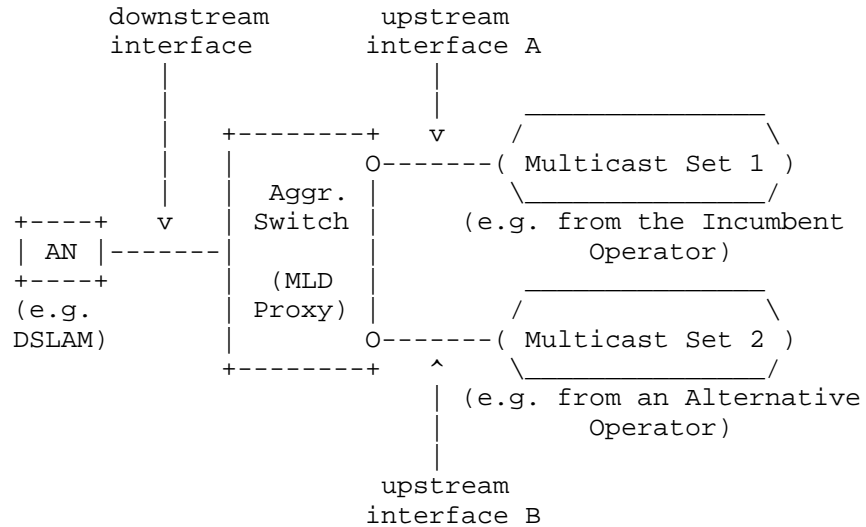


Figure 2: Example of usage of an MLD proxy with multiple upstream interfaces in a fixed network scenario

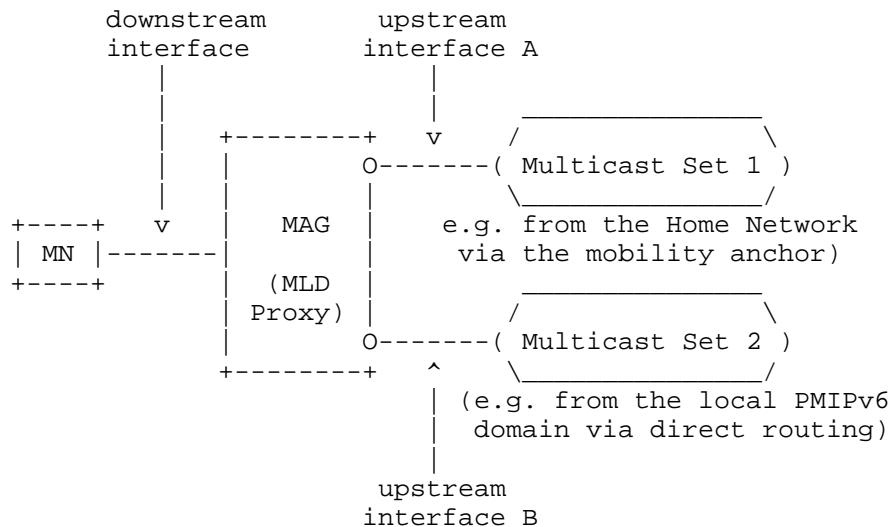


Figure 3: Example of usage of an MLD proxy with multiple upstream interfaces in a mobile network scenario

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

4. Scenarios of applicability

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

4.1. Fixed network scenarios

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

For the multicast service, the use of an MLD proxy with multiple

upstream interfaces in those switches can provide service flexibility in a lightweight and simpler manner if compared with PIM-routing based alternatives.

4.1.1. Multicast wholesale offer for residential services

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

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

4.1.1.1. Requirements

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

4.1.2. Multicast resiliency

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

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

4.1.2.1. Requirements

- o The MLD proxy should be able to deliver multicast control messages sent by the end user to the corresponding active upstream interface.

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

4.1.3. Load balancing for multicast traffic in the metro segment

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

4.1.3.1. Requirements

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

4.1.4. Summary of the requirements needed for mobile network scenarios

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

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

Figure 4: Functionality needed on MLD proxy with multiple upstream interfaces per application scenario in fixed networks

4.2. Mobile network scenarios

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

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

4.2.1. Applicability to multicast listener mobility

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

4.2.1.1. Single MLD proxy instance on MAG

The base solution for multicast service in PMIPv6 RFC6224 [RFC6224] assumes that any MN subscribed to multicast services receive the multicast traffic through the associated LMA, as in the unicast case. As standard MLD proxy functionality only supports one upstream interface, the MAG should implement several separated MLD proxy instances, one per LMA, in order to serve the multicast traffic to the MNs, according to any particular LMA-MN association.

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

4.2.1.1.1. Requirements

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

4.2.1.2. Remote and local multicast subscription

This scenario has been already introduced in the previous section, and can be seen in Figure 3. Standard MLD proxy definition, with a unique upstream interface per proxy, does not allow the reception of multicast traffic from distinct upstream multicast routers. In other words, all the multicast traffic being sent to the MLD proxy in downstream traverses a concrete, unique router before reaching the MAG. There are, however, situations where different multicast content could reach the MLD proxy through distinct next-hop routers.

For instance, the solution adopted to avoid the tunnel convergence problem in basic multicast PMIPv6 deployments [I-D.ietf-multimob-pmipv6-ropt] considers the possibility of subscription to a multicast source local to the PMIPv6 domain. In that situation, some multicast content will be accesses remotely, through the home network via the multicast tree mobility anchor,

while some other multicast content will reach the proxy directly, via a local router in the domain.

4.2.1.2.1. Requirements

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

4.2.1.3. Dual subscription to multicast groups during handover

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

A generic solution is being defined in [I-D.ietf-multimob-handover-optimization] to speed up the knowledge of the ongoing subscription by the nMAG. However, for the particular case that the underlying radio access technology supports layer-2 triggers (thus requiring extra capabilities on the mobile node), there could be inter-MAG cooperation for handover support if pMAG and nMAG are known in advance.

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

being established.

A similar approach is followed in [I-D.schmidt-multimob-fmipv6-pfmipv6-multicast] despite the solution proposed there differs from this approach (i.e., there is no consideration of an MLD proxy with multiple interfaces).

4.2.1.3.1. Requirements

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

4.2.2. Applicability to multicast source mobility

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

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

In the basic case of source mobility, the multicast source is connected to one of the downstream interfaces of an MLD proxy. According to the standard specification RFC4605 [RFC4605] every packet sent by the multicast source will be forwarded towards the root of the multicast tree.

However, linked to the mobility listener problem, there could be the case of simultaneous remote subscribers, subscribing to the multicast content through the home network, and local subscribers, requesting

the contents directly via a multicast router residing on the same PMIPv6 domain where the source is attached to.

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

Basic source mobility is being defined in RFC4605 [RFC4605] but the solution proposed there does not allow simultaneous co-existence of remote and local subscribers (i.e., the content sent by the source is either distributed locally to a multicast router in the PMIPv6 domain, or remotely by using the bi-directional tunnel towards the mobility anchor, but not both simultaneously).

4.2.2.1.1. Requirements

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

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

In a certain PMIPv6 domain can be MNs associated to distinct LMAs using the same MAG to get access to their corresponding home networks. For multicast communication, according to the base solution RFC6224 [RFC6224], each MN - LMA association implies a distinct MLD proxy instance to be invoked in the MAG.

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

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

4.2.2.2.1. Requirements

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

4.2.2.3. Route optimization support in source mobility for remote subscribers

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

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

A similar route optimization approach is proposed in [I-D.liu-multimob-pmipv6-multicast-ro].

4.2.2.3.1. Requirements

- o The MLD proxy should be able of forwarding (replicating) the multicast content to both kinds of upstream interfaces, inter-MAG tunnel interfaces and MAG to mobility anchor tunnel interface.
- o The MLD proxy should be able of handling control information incoming through any of the two types of upstream interfaces,

providing the expected behavior for each of the multicast trees (e.g., no forwarding traffic on one inter-MAG link once there are not more listeners requesting the content).

- o The MLD proxy should be able of routing the multicast data towards different upstream interfaces for both remote and route optimized subscriptions that could happen simultaneously.
- o The MLD proxy should be able of maintaining a 1:N association between an MN and both the remote and local MAGs (or downstream to upstream).

4.2.3. Summary of the requirements needed for mobile network scenarios

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

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

Figure 5: Functionality needed on MLD proxy with multiple upstream

interfaces per application scenario in mobile networks

5. Functional specification of an MLD proxy with multiple interfaces

To be completed

6. Security Considerations

To be completed

7. IANA Considerations

To be completed

8. Acknowledgments

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

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

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4605] Fenner, B., He, H., Haberman, B., and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.
- [RFC5213] Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "Proxy Mobile IPv6", RFC 5213, August 2008.

9.2. Informative References

- [I-D.ietf-multimob-handover-optimization]
Contreras, L., Bernardos, C., and I. Soto, "PMIPv6 multicast handover optimization by the Subscription Information Acquisition through the LMA (SIAL)", draft-ietf-multimob-handover-optimization-02 (work in progress), February 2013.
- [I-D.ietf-multimob-pmipv6-ropt]
Zuniga, J., Contreras, L., Bernardos, C., Jeon, S., and Y. Kim, "Multicast Mobility Routing Optimizations for Proxy Mobile IPv6", draft-ietf-multimob-pmipv6-ropt-06 (work in progress), June 2013.
- [I-D.liu-multimob-pmipv6-multicast-ro]
Liu, J. and W. Luo, "Routes Optimization for Multicast Sender in Proxy Mobile IPv6 Domain", draft-liu-multimob-pmipv6-multicast-ro-02 (work in progress), July 2012.
- [I-D.schmidt-multimob-fmipv6-pfmipv6-multicast]
Schmidt, T., Waehlich, M., Koodli, R., and G. Fairhurst, "Multicast Listener Extensions for MIPv6 and PMIPv6 Fast Handovers", draft-schmidt-multimob-fmipv6-pfmipv6-multicast-07 (work in progress), October 2012.
- [RFC6224] Schmidt, T., Waehlich, M., and S. Krishnan, "Base Deployment for Multicast Listener Support in Proxy Mobile IPv6 (PMIPv6) Domains", RFC 6224, April 2011.

Appendix A. Basic support for multicast listener with PMIPv6

This section briefly summarizes the operation of Proxy Mobile IPv6 RFC5213 [RFC5213] and how multicast listener support works with PMIPv6 as specified in RFC6224 [RFC6224].

Proxy Mobile IPv6 (PMIPv6) RFC5213 [RFC5213] is a network-based mobility management protocol which enables the network to provide mobility support to standard IP terminals residing in the network. These terminals enjoy this mobility service without being required to implement any mobility-specific IP operations. Namely, PMIPv6 is one of the mechanisms adopted by the 3GPP to support the mobility management of non-3GPP terminals in future Evolved Packet System (EPS) networks.

PMIPv6 allows a Media Access Gateway (MAG) to establish a distinct bi-directional tunnel with different Local Mobility Anchors (LMAs), being each tunnel shared by the attached Mobile Nodes (MNs). Each mobile node is associated with a corresponding LMA, which keeps track of its current location, that is, the MAG where the mobile node is attached. IP-in-IP encapsulation is used within the tunnel to forward traffic between the LMA and the MAG. Figure 4 (taken from RFC5213 [RFC5213]) shows the architecture of a PMIPv6 domain.

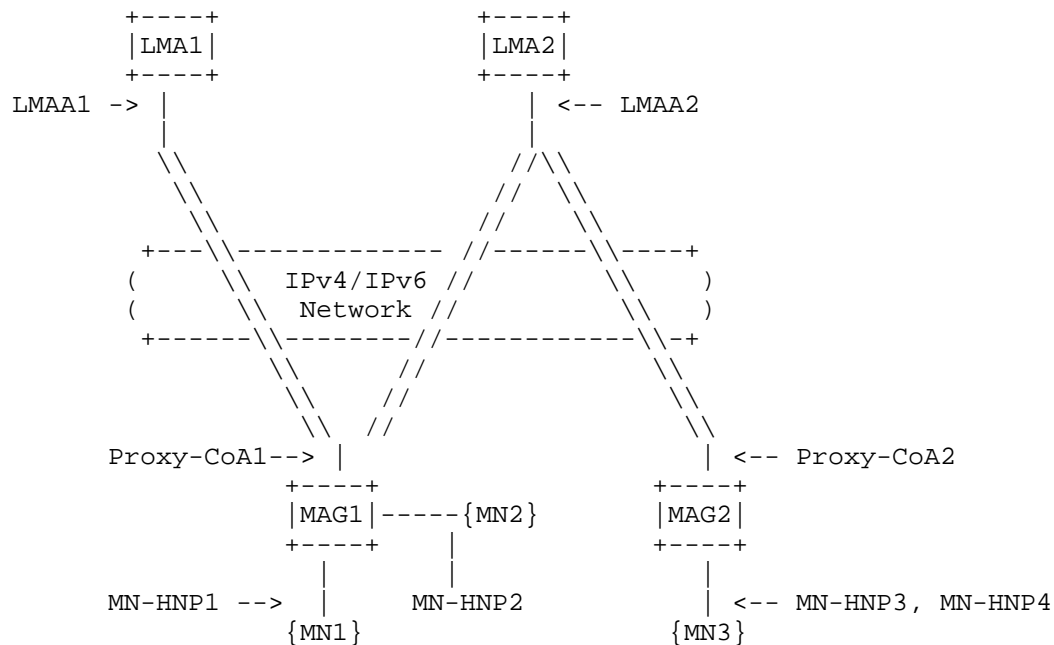


Figure 6: Proxy Mobile IPv6 Domain

The basic solution for the distribution of multicast traffic within a PMIPv6 domain RFC6224 [RFC6224] makes use of the bi-directional LMA-MAG tunnels. The base solution follows the so-called remote subscription model, in which the subscribed multicast content is delivered from the Home Network. By doing so, an individual copy of every multicast flow is delivered through the tunnel connecting the mobility anchor to any of the access gateways in the domain. In many cases, these individual copies traverse the same routers in the path towards the access gateways, incurring in an inefficient distribution, equivalent to the unicast distribution of the multicast content in the domain.

The reference scenario for multicast deployment in Proxy Mobile IPv6 domains is illustrated in Figure 5 (taken from RFC6224 [RFC6224]).

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

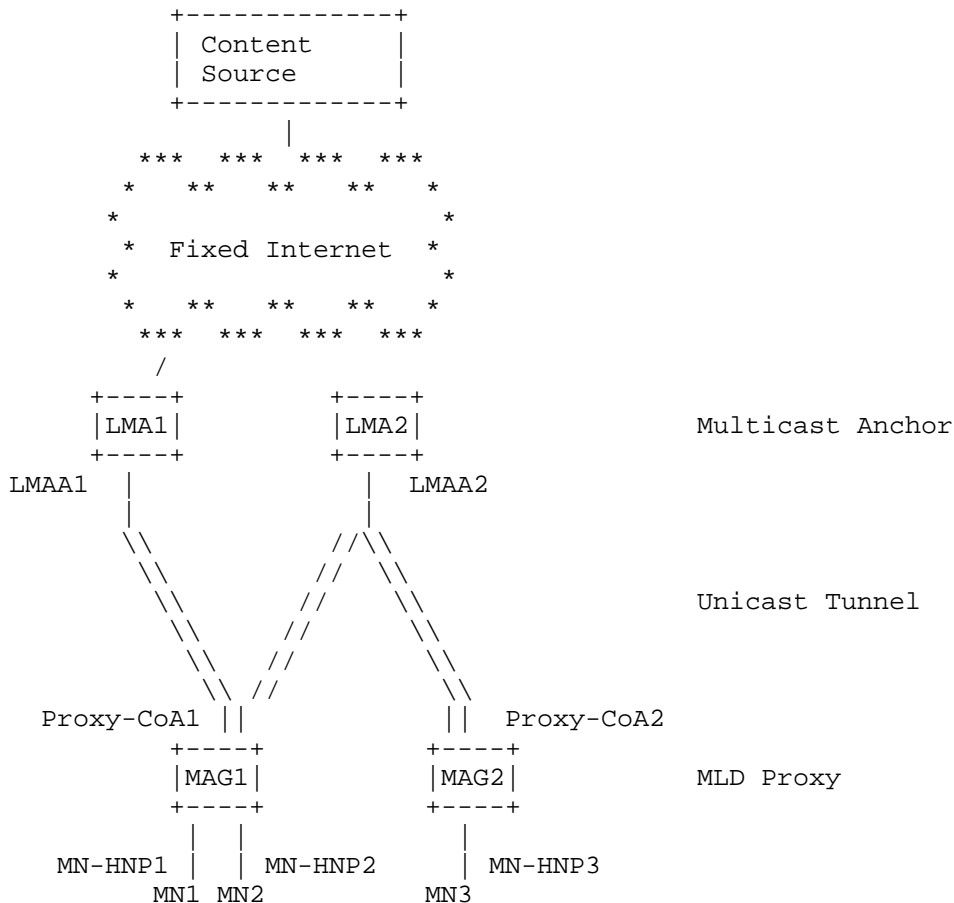


Figure 7: Reference Network for Multicast Deployment in PMIPv6

Authors' Addresses

Luis M. Contreras
Telefonica I+D
Ronda de la Comunicacion, s/n
Sur-3 building, 3rd floor
Madrid 28050
Spain

Email: lmcm@tid.es

Carlos J. Bernardos
Universidad Carlos III de Madrid
Av. Universidad, 30
Leganes, Madrid 28911
Spain

Phone: +34 91624 6236
Email: cjbc@it.uc3m.es
URI: <http://www.it.uc3m.es/cjbc/>

Juan Carlos
InterDigital Communications, LLC
1000 Sherbrooke Street West, 10th floor
Montreal, Quebec H3A 3G4
Canada

Email: JuanCarlos.Zuniga@InterDigital.com

6man Working Group
Internet-Draft
Updates: 3306,3956,4607,4291 (if approved)
Intended status: Standards Track
Expires: November 24, 2013

M. Boucadair
France Telecom
S. Venaas
Cisco
May 23, 2013

Updates to the IPv6 Multicast Addressing Architecture
draft-ietf-6man-multicast-addr-arch-update-01

Abstract

This document updates the IPv6 multicast addressing architecture by defining the 17-20 reserved bits as generic flag bits. The document provides also some clarifications related to the use of these flag bits.

This document updates RFC 3956, RFC 3306, RFC 4607 and RFC 4291.

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 RFC 2119 [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 November 24, 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	2
2. Addressing Architecture Update	2
3. Clarifications	3
3.1. Flag Bits	3
3.2. IANA Assigned SSM Block	4
4. RFC Updates	4
4.1. RFC3306	4
4.2. RFC3956	6
4.3. RFC4607	8
5. IANA Considerations	9
6. Security Considerations	9
7. Acknowledgements	9
8. Normative References	9
Authors' Addresses	10

1. Introduction

This document updates the IPv6 multicast addressing architecture [RFC4291] by defining the 17-20 reserved bits as generic flag bits (Section 2). The document provides also some clarifications related to the use of these flag bits (Section 3.1) and also about IANA assigned SSM blocks (Section 3.2).

This document updates [RFC3956], [RFC3306], [RFC4607] and [RFC4291].

2. Addressing Architecture Update

Bits 17-20 of a multicast address are defined in [RFC3956] and [RFC3306] as reserved bits. This document defines these bits as generic flag bits so that they apply to any multicast address. Figure 1 and Figure 2 show the updated structure of the addressing architecture. The first diagram shows the update of the base IPv6 addressing architecture, and the second shows the update of so-called Embedded-RP.

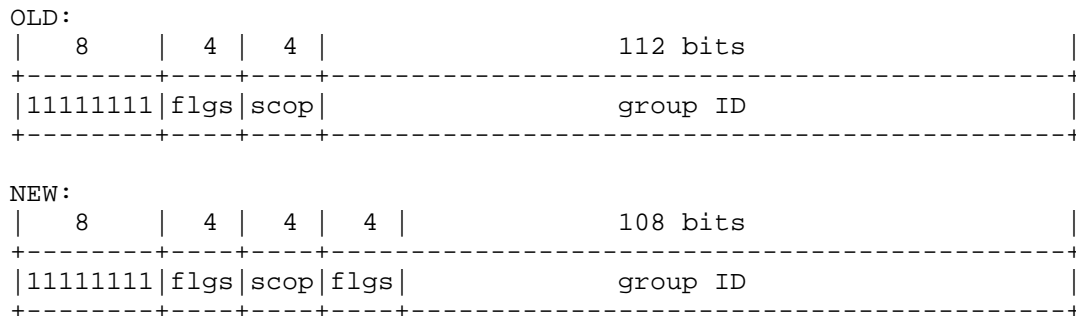


Figure 1: Updated IPv6 Multicast Addressing Architecture

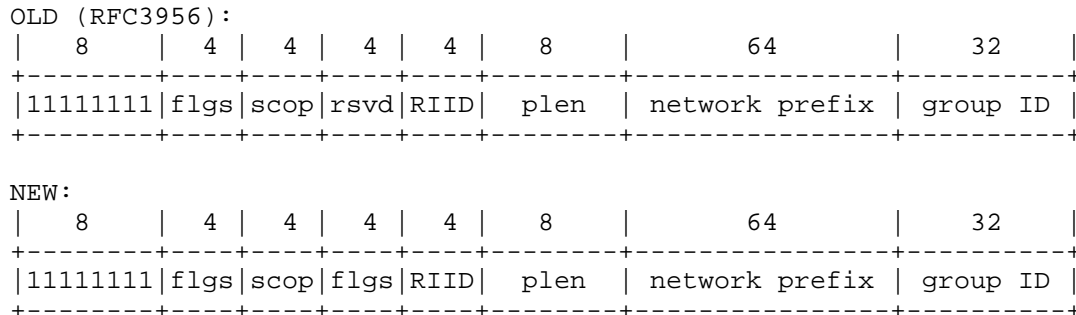


Figure 2: Embedded-RP with Updated IPv6 Multicast Address Arch.

Further specification documents may define a meaning for these flag bits. Defining the bits 17-20 as flags for all IPv6 multicast addresses allows addresses to be treated in a more uniform and generic way, and allows for these bits to be defined in the future for different purposes, irrespective of the specific type of multicast address.

3. Clarifications

3.1. Flag Bits

Some implementations and specification documents do not treat the flag bits as separate bits but tend to use their combined value as a 4-bit integer. This practice is a hurdle for assigning a meaning to the remaining flag bits. Below are listed some examples for illustration purposes:

- o the reading of [RFC4607] may lead to conclude that ff3x::/32 is the only allowed SSM IPv6 prefix block.

- o [RFC3956] states only ff70::/12 applies to Embedded-RP. Particularly, implementations should not treat the fff0::/12 range as Embedded-RP.

To avoid such confusion and to unambiguously associate a meaning with the remaining flags, the following recommendation is made

Implementations MUST treat flag bits as separate bits.

3.2. IANA Assigned SSM Block

Another issue related to SSM is the IANA assigned SSM address block. Per [RFC4607], ff3x::4000:0001 through ff3x::7fff:fff is the block for IANA assignments (<http://www.iana.org/assignments/ipv6-multicast-addresses/ipv6-multicast-addresses.xml>). However, IANA assignments are permanent addresses and should not have the transient bit set. Quoting from [RFC4607]:

"T = 1 indicates a non-permanently-assigned ("transient") multicast address."

4. RFC Updates

4.1. RFC3306

This document changes Section 4 of [RFC3306] as follows:

OLD:

8	4	4	8	8	64	32	
+-----+	+-----+	+-----+	+-----+	+-----+	+-----+	+-----+	+-----+
11111111	flgs	scop	reserved	plen	network prefix	group ID	
+-----+	+-----+	+-----+	+-----+	+-----+	+-----+	+-----+	+-----+

flgs is a set of 4 flags:

+--+--+--+
0 0 P T
+--+--+--+

- o P = 0 indicates a multicast address that is not assigned based on the network prefix. This indicates a multicast address as defined in [ADDRARCH].
- o P = 1 indicates a multicast address that is assigned based on the network prefix.
- o If P = 1, T MUST be set to 1, otherwise the setting of the T bit is defined in Section 2.7 of [ADDRARCH].

The reserved field MUST be zero.

NEW:

	8		4		4		8		8		64		32	
+-----+		+-----+		+-----+		+-----+		+-----+		+-----+		+-----+		+-----+
	11111111		flgs		scop		reserved		plen		network prefix		group ID	
+-----+		+-----+		+-----+		+-----+		+-----+		+-----+		+-----+		+-----+

flgs is a set of 4 flags: +--+--+--+
 |X|Y|P|T|
 +--+--+--+

X and Y may each be set to 0 or 1.

- o P = 0 indicates a multicast address that is not assigned based on the network prefix. This indicates a multicast address as defined in [ADDRARCH].
- o P = 1 indicates a multicast address that is assigned based on the network prefix.
- o T is set according to the definition in Section 2.7 of [ADDRARCH]. Unicast-Prefix-based addresses would typically not be IANA assigned, so in most cases T would be set to 1.

This document changes Section 6 of [RFC3306] as follows:

OLD:

These settings create an SSM range of FF3x::/32 (where 'x' is any valid scope value). The source address field in the IPv6 header identifies the owner of the multicast address.

NEW:

T flag is set according to whether the addresses are assigned by IANA.

If the flag bits are to 0011, these settings create an SSM range of ff3x::/32 (where 'x' is any valid scope value). The source address field in the IPv6 header identifies the owner of the multicast address. ff3x::/32 is not the only allowed SSM prefix range. For example, ff2x::/32 would be IANA assigned SSM addresses.

4.2. RFC3956

This document changes Section 2 of [RFC3956] as follows:

OLD:

As described in [RFC3306], the multicast address format is as follows:

	8		4		4		8		8		64		32	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+
	11111111		flgs		scop		reserved		plen		network prefix		group ID	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+

Where flgs are "0011". (The first two bits are as yet undefined, sent as zero and ignored on receipt.)

NEW:

As described in [RFC3306], the multicast address format is as follows:

	8		4		4		4		4		8		64		32	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+
	11111111		flgs		scop		flgs		rsvd		plen		network prefix		group ID	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+

flgs is a set of four flags:

+	-	+	-	+	-	+	-	+
	X		R		P		T	
+	-	+	-	+	-	+	-	+

X may be set to 0 or 1.

This document changes Section 3 of [RFC3956] as follows:

OLD:

	8		4		4		4		4		8		64		32	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+

```

|11111111|flgs|scop|rsvd|RIID|plen| network prefix | group ID |
+-----+-----+-----+-----+-----+-----+-----+
                                     +-+--+--+
flgs is a set of four flags:      |0|R|P|T|
                                     +--+--+--+

```

When the highest-order bit is 0, R = 1 indicates a multicast address that embeds the address on the RP. Then P MUST be set to 1, and consequently T MUST be set to 1, as specified in [RFC3306]. In effect, this implies the prefix FF70::/12. In this case, the last 4 bits of the previously reserved field are interpreted as embedding the RP interface ID, as specified in this memo.

The behavior is unspecified if P or T is not set to 1, as then the prefix would not be FF70::/12. Likewise, the encoding and the protocol mode used when the two high-order bits in "flgs" are set to 11 ("FFF0::/12") is intentionally unspecified until such time that the highest-order bit is defined. Without further IETF specification, implementations SHOULD NOT treat the FFF0::/12 range as Embedded-RP.

NEW:

```

| 8 | 4 | 4 | 4 | 4 | 8 | 64 | 32 |
+---+---+---+---+---+---+---+---+
|11111111|flgs|scop|flgs|RIID|plen| network prefix | group ID |
+---+---+---+---+---+---+---+
                                     +-+--+--+
flgs is a set of four flags:      |X|R|P|T|
                                     +--+--+--+

```

X may be set to 0 or 1.

R = 1 indicates a multicast address that embeds the address of the RP. P MUST be set to 1 according to [RFC3306], as this is a special case of unicast-prefix based addresses. This implies that for instance prefixes ff70::/12 and fff0::/12 are embedded RP prefixes, but all multicast addresses with the R-bit set to 1 MUST be treated as Embedded RP addresses. The behavior is unspecified if P is not set to 1. When the R-bit is set, the last 4 bits of the previously reserved field are interpreted as embedding the RP interface ID, as specified in this memo.

This document changes Section 4 of [RFC3956] as follows:

OLD:

It MUST be a multicast address with "flgs" set to 0111, that is, to be of the prefix FF70::/12,

NEW:

It MUST be a multicast address with R-bit set to 1.

It MUST have P-bit set to 1 when using the embedding in this document as it is a prefix-based address.

This document changes Section 7.1 of [RFC3956] as follows:

OLD:

To avoid loops and inconsistencies, for addresses in the range FF70::/12, the Embedded-RP mapping MUST be considered the longest possible match and higher priority than any other mechanism.

NEW:

To avoid loops and inconsistencies, for addresses with R-bit set to 1, the Embedded-RP mapping MUST be considered the longest possible match and higher priority than any other mechanism.

4.3. RFC4607

This document changes the abstract of [RFC4607] as follows:

OLD:

IP version 4 (IPv4) addresses in the 232/8 (232.0.0.0 to 232.255.255.255) range are designated as source-specific multicast (SSM) destination addresses and are reserved for use by source-specific applications and protocols. For IP version 6 (IPv6), the address prefix FF3x::/32 is currently reserved for source-specific multicast use but others may be reserved in the future. This document defines an extension to the Internet network service that applies to datagrams sent to SSM addresses and defines the host and router requirements to support this extension.

NEW:

IP version 4 (IPv4) addresses in the 232/8 (232.0.0.0 to 232.255.255.255) range are designated as source-specific multicast (SSM) destination addresses and are reserved for use by source-specific applications and protocols. For IP version 6 (IPv6), the address prefix ff3x::/32 is currently reserved for source-specific multicast use but others may be reserved in the future. This

document defines an extension to the Internet network service that applies to datagrams sent to SSM addresses and defines the host and router requirements to support this extension.

This document changes Section 1 of [RFC4607] as follows:

OLD:

For IPv6, the address prefix FF3x::/32 is reserved for source-specific multicast use, where 'x' is any valid scope identifier, by [IPv6-UBM]. Using the terminology of [IPv6-UBM], all SSM addresses must have P=1, T=1, and plen=0. [IPv6-MALLOC] mandates that the network prefix field of an SSM address also be set to zero, hence all SSM addresses fall in the FF3x::/96 range. Future documents may allow a non-zero network prefix field if, for instance, a new IP- address-to-MAC-address mapping is defined. Thus, address allocation should occur within the FF3x::/96 range, but a system should treat all of FF3x::/32 as SSM addresses, to allow for compatibility with possible future uses of the network prefix field.

NEW:

For IPv6, all SSM addresses must have P=1 and plen=0 while T-bit is set according to whether the addresses are assigned by IANA [I-D.ietf-6man-multicast-addr-arch-update]. In particular, a system should treat all of ff3x::/32 and ff2x::/32 as SSM addresses, to allow for compatibility with possible future uses of the network prefix field. Other SSM prefixes can be defined in the future.

5. IANA Considerations

This document may require IANA updates. However, at this point it is not clear exactly what these updates may be.

6. Security Considerations

Security considerations discussed in [RFC3956], [RFC3306], [RFC4607] and [RFC4291] MUST be taken into account.

7. Acknowledgements

Many thanks to B. Haberman for the discussions prior to the publication of this document.

8. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3306] Haberman, B. and D. Thaler, "Unicast-Prefix-based IPv6 Multicast Addresses", RFC 3306, August 2002.
- [RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, November 2004.
- [RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.
- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.

Authors' Addresses

Mohamed Boucadair
France Telecom
Rennes 35000
France

Email: mohamed.boucadair@orange.com

Stig Venaas
Cisco
USA

Email: stig@cisco.com

PIM
Internet-Draft
Intended status: Experimental
Expires: January 10, 2014

B. Joshi
J. Kaveetil
Infosys Ltd.
July 9, 2013

PIM neighbor selection with ECMP routes
draft-joshi-pim-ecmp-neighbor-select-00

Abstract

A Protocol Independent Multicast (PIM) router uses local forwarding table to select the upstream PIM neighbor towards the source or Rendezvous Point (RP). A router need to choose one upstream PIM neighbor from the list of PIM neighbors if the route for a source or RP is an Equal Cost Multipath (ECMP) route. Currently PIM routers use vendor specific algorithm to choose one neighbor. This may lead to unnecessary wastage of network resources. This draft first explains the need for a standard algorithm to select a PIM neighbor with ECMP route and then suggest such an algorithm.

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 RFC 2119 [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 January 10, 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	3
2. Need for a standard algorithm	3
3. Proposed algorithm	4
4. Security Considerations	5
5. Acknowledgements	5
6. References	5
6.1. Normative References	5
6.2. Informative References	5
Authors' Addresses	5

1. Introduction

A PIM router uses local forwarding table to select the upstream PIM neighbor towards the source or RP. A router need to choose one upstream PIM neighbor from a list of PIM neighbors if the route for a source or RP is an ECMP route. Currently PIM routers use some vendor specific algorithm to choose one PIM neighbor from the list of PIM neighbors. This may lead to wastage of network resources if two downstream routers on a LAN chooses two different upstream routers to reach a source or RP.

There is a need for routers to use a standard algorithm to choose the same PIM neighbor when an ECMP routes provide a list of PIM neighbors. This draft first explains the usefulness of such a standard algorithm and then suggest the same.

2. Need for a standard algorithm

Let us look at the network configuration of figure 1. Let us assume that R4 and R5 received a PIM join for a source 'S'. Let us also assume that both R4 and R5 has an ECMP route to reach source 'S'. Let us assume that this ECMP route has two gateways R1 and R3.

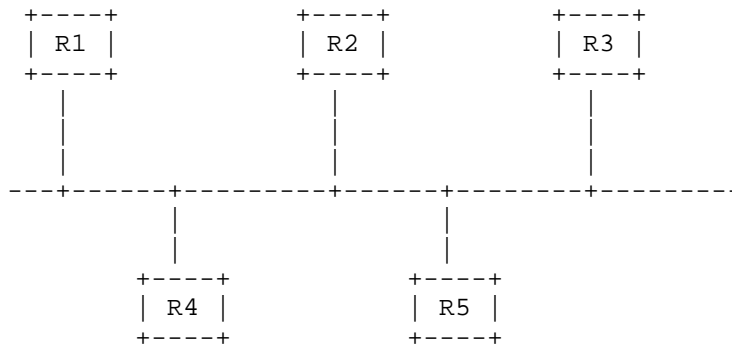


Figure 1

There is no guarantee that R4 and R5 will select the same upstream router i.e. R1 or R3 for the PIM join. This may be because R4 and R5 are from two different vendors or they use two different mechanisms to select a PIM neighbor. Let us assume R4 selects R1 while R5 selects R3 as the upstream PIM neighbor and sends the PIM join to them. R1 and R3 will forward these joins towards the source and somewhere these joins may or may not converge in the network. When both R1 and R3 forward the multicast traffic, PIM assert techniques will elect one

of them as forwarder.

So selecting two different PIM neighbors lead to following:

- o A complete leg of multicast tree will be unnecessarily created.
- o Multicast data will be unnecessarily forwarded along this leg.
- o An unnecessary assert mechanism will trigger in the LAN.

3. Proposed algorithm

These wastages can be avoided, if router R4 and R5 uses some standard algorithm to determine the upstream PIM neighbor. Any algorithm will work only if the routing protocols have converged and all routers on a LAN have the same forwarding table. One simple mechanism could be to choose the highest IP address PIM neighbor from the list of PIM neighbors. However, with such a mechanism, all multicast channels using a specific ECMP route, would end up using the same PIM neighbor. This means that there is no load balancing among the PIM neighbors in the list. So this draft proposes to use a hash algorithm which uses PIM neighbor's address, multicast group address and multicast source address or RP address. The hash function should implicitly provide the load balancing among the PIM neighbors that can be used.

The hash function explained in PIM-SM [RFC4601] is extended. The hash function proposed here will be used to calculate hash value for each combination of source/RP, group and PIM neighbor's primary address. The PIM neighbor with the highest value will be chosen as the upstream PIM neighbor for the corresponding multicast group and source.

```
HashValue_for_ASM(Group-address,RP-address,Neighbor-address) =  
  ( 1103515245 *  
    ( ( 1103515245 *  
      ( ( ( 1103515245 * Group-address ) + 12345 )  
        XOR ( RP-address ) ) + 12345 )  
      XOR ( Neighbor-address ) ) + 12345 ) mod 2^31
```

```
HashValue_for_SSM(Group-address,Source-address,Neighbor-address) =  
  ( 1103515245 *  
    ( ( 1103515245 *  
      ( ( ( 1103515245 * Group-address ) + 12345 )  
        XOR ( Source-address ) ) + 12345 )  
      XOR ( Neighbor-address ) ) + 12345 ) mod 2^31
```

4. Security Considerations

This draft does not suggest any change in protocol messages. If all routers in a LAN support this document but one router choose different upstream router than others, the situation will be similar to what exists today.

5. Acknowledgements

The idea to use a hash function to identify the PIM neighbor in the list of PIM neighbors was discussed during an informal discussion with Saravana Prasad.

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

6.2. Informative References

Authors' Addresses

Bharat Joshi
Infosys Ltd.

Email: bharat_joshi@infosys.com

Jithesh Kaveetil
Infosys Ltd.

Email: jithesh_k@infosys.com

Protocol Independent Multicast Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2014

R. Kebler, Ed.
A. Atlas
Juniper Networks
N. Shen
Cisco Systems, Inc.
Y. Cai
Microsoft
July 12, 2013

PIM Extensions for Protection Using Maximally Redundant Trees
draft-kebler-pim-mrt-protection-01

Abstract

This document specifies Protocol Independent Multicast (PIM) procedures for Failure Protection, as specified in the MRT Multicast architecture [I-D.atlas-rtwg-mrt-mc-arch]. This can be accomplished with Global Repair (aka Live-Live) or with Local Repair (aka Fast Re-route). Maximally Redundant Trees (MRTs) provide the capability to PIM to provide alternate paths around any given failure.

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 January 13, 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. Terminology	3
3. Global Protection (Live-Live)	4
3.1. Egress Router Behavior	4
3.2. Limitation when a LAN is a cut-link	4
3.3. Using Different Groups to identify MRTs	5
4. Local Protection	5
4.1. PLR Replication	5
4.1.1. PLR Behavior	6
4.1.2. Unicast convergence during PLR Replication	6
4.1.3. MP Behavior	6
4.1.4. Downstream Routers from the MP	7
4.1.5. Protected Node Behavior	7
5. Packet Formats	8
5.1. Hello Options	8
5.1.1. MRT Protection Capabilities	8
5.2. Join Attributes	9
5.2.1. Merge Point Attribute	9
6. IANA Considerations	10
7. Security Considerations	10
8. References	10
8.1. Normative References	10
8.2. Informative References	11
Authors' Addresses	11

1. Introduction

This document specifies how to reduce traffic loss after network failures by using Maximally Redundant Trees (MRTs). This can be accomplished with Global Repair (aka Live-Live) or with Local Repair (aka Fast Re-route). The tradeoffs and applicability for each method of protection are discussed in the MRT Multicast architecture [I-D.atlas-rtwg-mrt-mc-arch].

With Global Repair, a multicast egress will send PIM Joins for the same stream on multiple MRT topologies. The Global Repair specified in this document is similar to [I-D.ietf-rtgwg-mofrr]. This document specifies how this can be accomplished using MRTs, however, providing 100% coverage without requiring any particular network topology.

Local Repair for Link or Node protection can also be used to protect Multicast traffic. A Point of Local Repair (PLR) can replicate the traffic to all Merge Points (MPs). In order to accomplish this, the PLR must know the unicast destination of all MPs. Upon the failure, the PLR will send the traffic to all MPs.

2. Terminology

2-connected: A graph that has no cut-vertices. This is a graph that requires two nodes to be removed before the network is partitioned.

cut-link: A link whose removal partitions the network. A cut-link by definition must be connected between two cut-vertices. If there are multiple parallel links, then they are referred to as cut-links in this document if removing the set of parallel links would partition the network.

cut-vertex: A vertex whose removal partitions the network.

Maximally Redundant Trees (MRT): A pair of trees where the path from any node X to the root R along the first tree and the path from the same node X to the root along the second tree share the minimum number of nodes and the minimum number of links. Each such shared node is a cut-vertex. Any shared links are cut-links. Any RT is an MRT but many MRTs are not RTs.

Maximally Redundant Multicast Trees (MRMT): A pair of multicast trees built of the sub-set of MRTs that is needed to reach all interested receivers.

network graph: A graph that reflects the network topology where all links connect exactly two nodes and broadcast links have been transformed into the standard pseudo-node representation.

Redundant Trees (RT): A pair of trees where the path from any node X to the root R along the first tree is node-disjoint with the path from the same node X to the root along the second tree. These can be computed in 2-connected graphs.

Merge Point (MP): For local repair, a router at which the alternate traffic rejoins the primary multicast tree. For global protection, a router which receives traffic on multiple trees and must decide which stream to forward on.

Point of Local Repair (PLR): The router that detects a local failure and decides whether and when to forward traffic on appropriate alternates.

MT-ID: Multi-topology identifier

Stream Selection: The process by which a router determines which of the multiple primary multicast streams to accept and forward. The router can decide on a packet-by-packet basis or simply per-stream. This is done for global protection as described in [I-D.ietf-rtgwg-mofrr].

MultiCast Egress (MCE): Multicast Egress, a node where the multicast stream exists the current PIM domain. This is usually a receiving router that may forward the multicast traffic towards receivers based upon IGMP or other technology.

3. Global Protection (Live-Live)

In order to achieve Global Protection, traffic will flow through the network through disjoint paths. A multicast egress (MCE) router will trigger this traffic flow by sending PIM joins on two different interfaces. The MRT algorithm will ensure that these joins travel through maximally disjoint paths to the source. The egress router must then forward a single stream along to its downstream members. Any failure in the network to either stream can be repaired by this egress router, since it is receiving the redundant stream.

Any router that is capable of supporting Global Repair with MRTs must advertise the T bit in the MRT Protection Hello Option.

3.1. Egress Router Behavior

A multicast egress (MCE) router will join a multicast stream on both the Blue and Red MRT. The MT-ID [RFC6420] of the MRT will be included in the Join as a Join Attribute [RFC5384]. Traffic will flow down both MRTs to the egress router to achieve redundancy. The egress router will forward a single stream along to its downstream interfaces. The techniques for this stream selection are described in MoFRR [I-D.ietf-rtgwg-mofrr].

3.2. Limitation when a LAN is a cut-link

There is a limitation in end-to-end protection when, for a given S,G, the MRTs converge on a single LAN with different upstream neighbors. In this case, both upstream neighbors will be sending on the LAN, and there is no distinguishing the data traffic for the different MRTs if it is carried with the same S,G. The PIM Assert procedures will select a single forwarding router on the LAN and the other router will stop sending. This could cause the Assert Loser to prune back the S,G. Therefore, traffic will flow on only one MRT between the source and the downstream router on the LAN.

3.3. Using Different Groups to identify MRTs

There may be cases when different sources or groups are used to send the same stream, as described in the MRT Multicast architecture [I-D.atlas-rtwg-mrt-mc-arch]. In this case the egress router may not need to perform the stream selection. However, it would be desirable for the egress router to join the sources or groups on different MRTS. The mechanism to perform group to MRT mapping is outside the scope of this document. Once the egress router knows the group to MRT mapping, then it will join for the S,G on the particular tree by including the MT-ID for the MRT in the Join message. In this case, the streams can travel across the same LAN without the issues described above.

4. Local Protection

Local Protection can protect either a link or node, and this will be determined on a per flow basis. A Join Attribute will be used for downstream routers to signal the Merge Point information. Each router will advertise in its MRT Protection Hello Options whether it is capable of performing Link or Node protection.

4.1. PLR Replication

At a PLR, each S,G flow will have a set of downstream interfaces and a set of MPs for each downstream interface. There will be MPLS label information learned for each MP. Upon a failure to the protected link, the PLR will encapsulate and send the protected multicast traffic to all MPs for that particular (S,G,intf). The MP will, therefore, receive the encapsulated data upon the failure and traffic will resume to all of its downstream receivers. Once the PLR has given the downstream routers sufficient time to recover from the failure, it can stop sending the protected traffic, and prune upstream, if required.

For the PLR to send the protected traffic upon a failure, it requires the unicast address and an MPLS label (which may be Implicit Null) for all the Merge Points. Each MP will advertise this information in a Merge Point Join Attribute. If link protection is used, this is sufficient to reach the PLR. For node protection, the information for all MPs will be sent to the PLR in a Join Attribute from the upstream node of the MP (i.e., the Protected Node). In this case, the MP will set the N bit in these Join Attributes to indicate the Protected Node needs to send the Join Attribute upstream to the PLR. If the MP or the Protecting Node is sending the Join attribute to the PLR, it will set the P bit in the Join Attribute.

All routers that support this functionality will advertise the Link or Node capability bits in the MRT Protection Hello Option. Any Node that is capable of acting as a PLR will advertise the PLR-Replication capable bit in the MRT Protection Hello Option.

4.1.1. PLR Behavior

The PLR will learn the location of all the MPs in the its Join Messages that it receives from downstream routers. The Merge Points will be kept per (S,G, downstream-interface). Upon a failure to the protected interface, the PLR will encapsulate and forward the multicast data to all the MPs for that downstream interface, and it will start the Alternate-Tree-Protection-Timer. The Alternate-Tree-Protection-Timer should be a configurable timer with a default of 10 seconds. The PLR will suppress the PIM Prunes from being sent while the Protection-Timer is running. Once this timer expires, it will stop sending the traffic to MPs, and it can send a Prune upstream if required.

For a PLR to learn of all MPs, then Join Suppression must be disabled on the interfaces between the MP and the PLR. In addition, the PLR must accept all MP ID Join Attributes that it receives from downstream neighbors.

4.1.2. Unicast convergence during PLR Replication

Since it is likely for unicast routes to converge before PIM fully converges, the PLR must still be able to route the traffic to all MPs while unicast recovers from the original failure. The PLR must not use stale forwarding information to reach the MPs for the protected multicast traffic if unicast has already updated its forwarding entries after the network event. An implementation should use the same forwarding information that would be used to forward unicast traffic to that destination. In this way, the PLR will be able to forward traffic to the MPs.

4.1.3. MP Behavior

As is done today, the MP will forward traffic received on its normal incoming interface. While the normal RPF interface is up, encapsulated alternate traffic will not be forwarded. If the RPF interface fails, the MP will forward the encapsulated alternate traffic (if it is received with the correct encapsulation). This procedure assumes that there is a method for the routers on both sides of the protected link to determine if the link has gone down. Such methods are outside the scope of this document.

After the incoming interface changes the MP will start the Alternate-Tree-Protection-Timer. Once traffic arrives on the new incoming interface or the Alternate-Tree-Protection-Timer expires, the Merge Point will advertise the label for the new RPF interface in the Merge Point Join Attribute, and it will stop accepting the encapsulated alternate traffic.

The MP needs to know when it can release the label that it has advertised and potentially re-use that label for another purpose. If the interface goes down or the adjacency goes down on an interface that the MP was advertising a label, it should wait JP_Holdtime for link protection and (2 * JP_Holdtime) for node protection before re-using that label for any other purpose.

4.1.4. Downstream Routers from the MP

Some make-before-break techniques should be used on routers downstream from the failure to ensure that traffic is not discarded once these routers learn of the unicast change. For example, if a downstream router, upon a unicast route change, prunes itself off its old RPF interface and discards traffic until the new tree is formed back to the source, then there will be end-to-end loss. The work that the upstream routers did to repair the local failure will be wasted since the downstream router is going to discard flowing traffic. The make-before-breaks procedures needed on the downstream router is outside the scope of this document.

4.1.5. Protected Node Behavior

For Node Protection, the MP will be one hop away from the Protected Node and two hops away from the PLR. In this case, there may be multiple next-next-hops to advertise as Merge Points in the Join Attribute. The Protected Node will learn the downstream members and it will gather the MP information from each downstream neighbor's Merge Point Join Attribute. For each Merge Point in the downstream list, the Protected Node will include a Merge Point Join Attribute in the Join that is sent upstream to the PLR. These Join attributes must have the N bit cleared when they are sent to the PLR. The PLR will add a Merge Point attribute for its own information to include itself as a Merge Point. All the Join attributes will have the P bit set, indicating they are being sent to a PLR. The Merge Point information may change for a route entry before the JoinPrune would normally be updated or refreshed to the PLR. Upon a change to the next-next hop list, the router can send a triggered JoinPrune with the updated Join Attribute, or it can wait for the next periodic refresh. It would be a tradeoff of increased control messages against a window of being unprotected. For a PLR to learn of all MPs, then Join Suppression must be disabled on the interfaces between the MP and the PLR.

5. Packet Formats

5.1. Hello Options

5.1.1. MRT Protection Capabilities

```

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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     |                                     |
|                                     Type = TBD                             |
|                                     |                                     |
|                                     Length = 1                             |
|                                     |                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Rsvd | P | T | L | N |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

MRT Protection Hello Option Format

Type: TBD.

Length: 1

Rsvd: Sent with 0, ignored on receipt

P: PLR Replication capable. This bit is set if a router is capable of acting as a PLR-replicating router, as described in this document. This router must also be capable of receiving a Merge Point Join Attribute.

T: MRT Topology Capable. This bit is set if the router is capable of understanding MRT topology IDs sent in the MT-ID Join Attribute [RFC5384], as defined in this document.

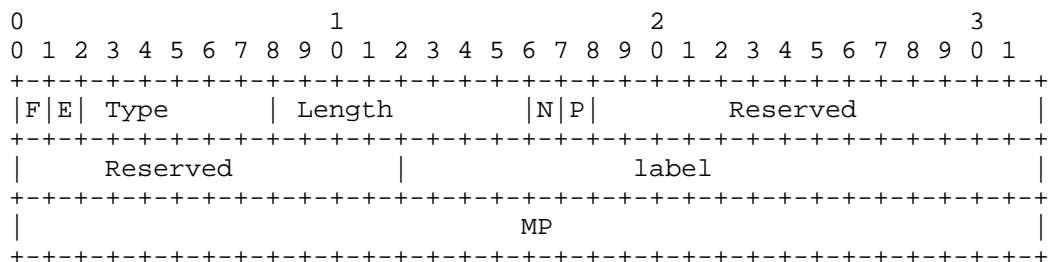
L: Link Protection Capable. This bit is set if the router is capable of performing Link Protection, as defined in this document. This router must also be capable of receiving a Merge Point Join Attribute.

N: Node Protection. This bit is set if the router is capable of performing Node Protection, as defined in this document. This router must also be capable of receiving a Merge Point Join Attribute.

5.2. Join Attributes

5.2.1. Merge Point Attribute

The following Join attribute is used for local protection, when the Protected-Node needs to signal the Merge Point information to the PLR. There will be a separate Merge Point Attribute for each Merge Point being advertised for the source. This attribute should only be sent to routers that are Link or Node capable, as advertised in the MRT Protection Hello Option.



Merge Point Join Attribute

F-bit: This bit will be clear as this is a non-transitive attribute.

E-bit: As defined in [RFC5384]

Type: TBD

Length field: variable

N bit - This Label is for a node-protecting MP. The label and this Join attribute will need to be sent upstream (to the PLR) in the upstream Join message. When sending this Join attribute upstream, this bit MUST be cleared.

P bit - This bit indicates that the receiving router should act as a PLR. This bit should only be set in Joins to routers that are PLR-Replication capable, as advertised in the MRT Protection Hello Option

Reserved: Sent with zero, ignored on receipt

label: the MPLS label associated with this MP

MP: The encoded-Unicast addresses of the Merge Point

6. IANA Considerations

A new PIM Hello Option type is requested to assign to the MRT Protection Hello Option

A new PIM Join Attribute Type is requested for the Merge Point Join Attribute

7. Security Considerations

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

8. References

8.1. Normative References

[I-D.ietf-rtgwg-mofrr]

Karan, A., Filsfils, C., Farinacci, D., Wijnands, I., Decraene, B., Joorde, U., and W. Henderickx, "Multicast only Fast Re-Route", draft-ietf-rtgwg-mofrr-02 (work in progress), June 2013.

[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.
- [RFC6395] Gulrajani, S. and S. Venaas, "An Interface Identifier (ID) Hello Option for PIM", RFC 6395, October 2011.
- [RFC6420] Cai, Y. and H. Ou, "PIM Multi-Topology ID (MT-ID) Join Attribute", RFC 6420, November 2011.

8.2. Informative References

- [I-D.atlas-rtwg-mrt-mc-arch]
Atlas, A., Kebler, R., Wijnands, IJ., and G. Enyedi, "An Architecture for Multicast Protection Using Maximally Redundant Trees", atlas-rtwg-mrt-mc-arch-02 (work in progress), July 2013.
- [I-D.ietf-rtgwg-mrt-frr-architecture]
Atlas, A., Kebler, R., Enyedi, G., Csaszar, A., Tantsura, J., Konstantynowicz, M., White, R., and M. Shand, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-02 (work in progress), February 2013.

Authors' Addresses

Robert Kebler (editor)
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: rkebler@juniper.net

Alia Atlas
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: akatlas@juniper.net

Naiming Shen
Cisco Systems, Inc.
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: naiming@cisco.com

Yiqun Cai
Microsoft
La Avenida
Mountain View, CA 94043
USA

Email: yiqunc@microsoft.com

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 13, 2014

IJ. Wijnands
S. Venaas
Cisco Systems, Inc.
M. Brig
Aegis BMD Program Office
July 12, 2013

PIM flooding mechanism and source discovery
draft-wijnands-pim-source-discovery-bsr-03

Abstract

PIM Sparse-Mode uses a Rendezvous Point (RP) and shared trees to forward multicast packets to Last Hop Routers (LHR). After the first packet is received by the LHR, the source of the multicast stream is learned and the Shortest Path Tree (SPT) can be joined. This draft proposes a solution to support PIM Sparse Mode (SM) without the need for PIM registers, RPs or shared trees. Multicast source information is flooded throughout the multicast domain using a new generic PIM flooding mechanism. This mechanism is defined in this document, and is modeled after the PIM Bootstrap Router protocol. By removing the need for RPs and shared trees, the PIM-SM procedures are simplified, improving router operations, management and making the protocol more robust.

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 January 13, 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. Conventions used in this document	3
1.2. Terminology	3
2. A generic PIM flooding mechanism	3
2.1. PFP message format	4
3. Distributing Source to Group Mappings	5
3.1. Group Source Holdtime TLV	5
4. Originating SG messages	6
5. Processing SG messages	6
6. The first packets and bursty sources	7
7. Resiliency to network partitioning	8
8. Security Considerations	8
9. IANA considerations	8
10. Acknowledgments	8
11. References	9
11.1. Normative References	9
11.2. Informative References	9
Authors' Addresses	9

1. Introduction

PIM Sparse-Mode uses a Rendezvous Point (RP) and shared trees to forward multicast packets to Last Hop Routers (LHR). After the first packet is received by the LHR, the source of the multicast stream is learned and the Shortest Path Tree (SPT) can be joined. This draft proposes a solution to support PIM Sparse Mode (SM) without the need for PIM registers, RPs or shared trees. Multicast source information is flooded throughout the multicast domain using a new generic PIM flooding mechanism. This mechanism is defined in this document, and is modeled after the Bootstrap Router protocol [RFC5059]. By removing the need for RPs and shared trees, the PIM-SM procedures are

simplified, improving router operations, management and making the protocol more robust.

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].

1.2. Terminology

RP: Rendezvous Point.

BSR: Bootstrap Router.

RPF: Reverse Path Forwarding.

SPT: Shortest Path Tree.

FHR: First Hop Router, directly connected to the Source.

LHR: Last Hop Router, directly connected to the receiver.

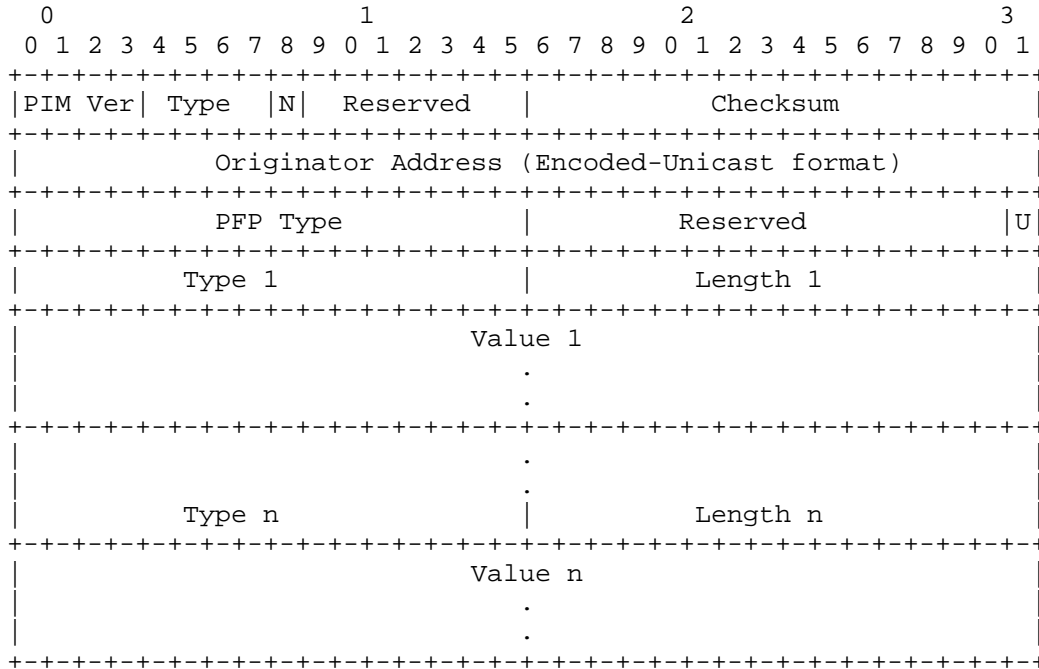
SG Mapping: Multicast source to group mapping.

SG Message: A PIM message containing SG Mappings.

2. A generic PIM flooding mechanism

The Bootstrap Router protocol (BSR) [RFC5059] is a commonly used protocol for distributing dynamic Group to RP mappings in PIM. It is responsible for flooding information about such mappings throughout a PIM domain, so that all routers in the domain can have the same information. BSR as defined, is only able to distribute Group to RP mappings. We are defining a more generic mechanism that can flood any kind of information throughout a PIM domain. It is not necessarily a domain though, it depends on administrative boundaries being configured. The forwarding rules are identical to BSR, except that there is no BSR election. The protocol includes an originator address which is used for RPF checking to restrict the flooding, just like BSR. Just like BSR it is also sent hop by hop. Note that there is no built in election mechanism as in BSR, so there can be multiple originators. It is still possible to add such an election mechanism if this protocol is used in scenarios where this is desirable. We include a type field, which can allow boundaries to be defined, and election to take place, independently per type. We call this protocol the PIM Flooding Protocol (PFP).

2.1. PFP message format



PIM Version: Reserved, Checksum Described in [RFC4601].

Type: PIM Message Type. Value (pending IANA) for a PFP message.

[N]o-Forward bit: When set, this bit means that the PFP message is not to be forwarded.

Originator Address: The address of the router that originated the message. This can be any address assigned to this router, but MUST be routable in the domain to allow successful forwarding (just like BSR address). The format for this address is given in the Encoded-Unicast address in [RFC4601].

PFP Type: There may be different sub protocols or different uses for this generic protocol. The PFP Type specifies which sub protocol it is used for.

[U]nknown-No-Forwarding bit: Some sub protocols may require each router to do some processing of the contents and not simply forwarding. This bit controls how a router should treat an unknown PFP Type. When set, a router MUST NOT forward the message

when the PFP Type is unknown. When clear, a router MUST forward the message when possible. If the PFP Type is known, then the specification of that type will specify how to handle the message, including whether it should be forwarded.

Type 1..n: A message contains one or more TLVs, in this case n TLVs. The Type specifies what kind of information is in the Value. Note that the Type space is shared between all PFP. Not all types make sense for all protocol types though.

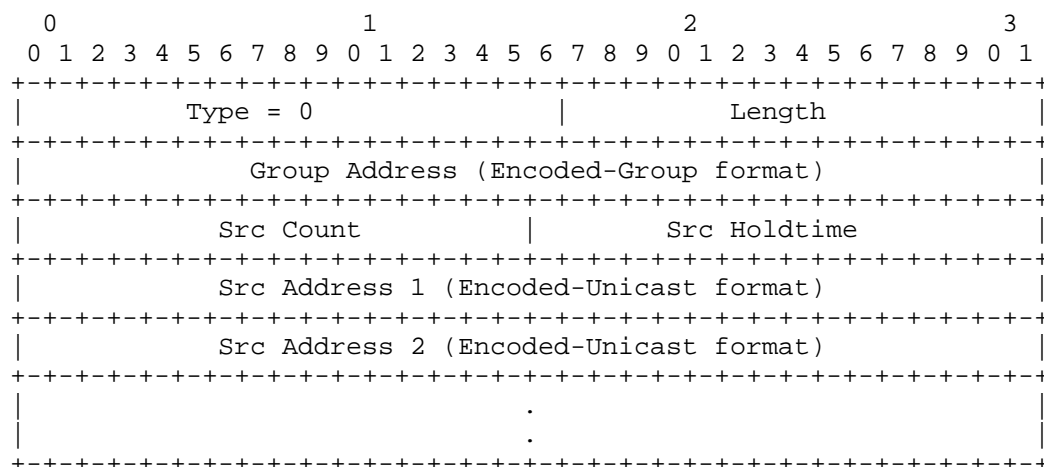
Length 1..n: The length of the the value field.

Value 1..n: The value associated with the type and of the specified length.

3. Distributing Source to Group Mappings

We want to provide information about active multicast sources throughout a PIM domain by making use of the generic flooding mechanism defined in the previous section. We request PFP Type 0 to be assigned for this purpose. We call a message with PFP Type 0 an SG Message. We also define a PFP TLV which we request to be type 0. How this TLV is used with PFP Type 0 is defined in the next section. Other PFP Types may specify the use of this TLV for other purposes. For PFP Type 0 the U-bit MUST NOT be set. This means that routers not supporting PFP Type 0 would still forward the message.

3.1. Group Source Holdtime TLV



```

|          Src Address m (Encoded-Unicast format)          |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Type: This TLV has type 0.

Length: The length of the value.

Group Address: The group we are announcing sources for. The format for this address is given in the Encoded-Group format in [RFC4601].

Src Count: How many unicast encoded sources address encodings follow.

Src Holdtime: The Holdtime (in seconds) for the corresponding source(s).

Src Address: The source address for the corresponding group. The format for these addresses is given in the Encoded-Unicast address in [RFC4601].

4. Originating SG messages

An SG Message, that is a PFP message of Type 0, may contain one or more Group Source Holdtime TLVs. This is used to flood information about active multicast sources. Each FHR that is directly connected to an active multicast source originates SG BSR messages. How a multicast router discovers the source of the multicast packet and when it considers itself the FHR follows the same procedures as the registering process described in [RFC4601]. After it is decided that a register needs to be sent, the SG is not registered via the PIM SM register procedures, but the SG mapping is included in an SG message. Note, only the SG mapping is distributed in the message, not the entire packet as would have been done with a PIM register. The router originating the SG messages includes one of its own addresses in the originator field. Note that this address must be routeable due to RPF checking. The SG messages are periodically sent for as long as the multicast source is active, similar to how PIM registers are periodically sent. The default announcement period is 60 seconds, which means that as long as the source is active, it is included in an SG message originated every 60 seconds. The holdtime for the source is by default 210 seconds. Other values can be configured, but the holdtime must be larger than the announcement period.

5. Processing SG messages

A router that receives an SG message should parse the message and store the SG mappings with a holdtimer started with the advertised holdtime for that group. If there are directly connected receivers for that group this router should send PIM (S,G) joins for all the SG mappings advertised in the message. The SG mappings are kept alive for as long as the holdtimer for the source is running. Once the holdtimer expires a PIM (S,G) prune must be sent to remove itself from the tree.

6. The first packets and bursty sources

The PIM register procedure is designed to deliver Multicast packets to the RP in the absence of a native SPT tree from the RP to the source. The register packets received on the RP are decapsulated and forwarded down the shared tree to the LHRs. As soon as an SPT tree is built, multicast packets would flow natively over the SPT to the RP or LHR and the register process would stop. The PIM register process bridges the gap between how long it takes to build the SPT tree to the FHR. If the packets would not be unicast encapsulated to the RP they would be dropped by the FHR until the SPT is setup. This functionality is important for applications where the initial packet(s) must be received for the application to work correctly. Another reason would be for bursty sources. If the application sends out a multicast packet every 4 minutes (or longer), the SPT is torn down (typically after 3:30 minutes of inactivity) before the next packet is forwarded down the tree. This will cause no multicast packet to ever be forwarded. A well behaved application should really be able to deal with packet loss since IP is a best effort based packet delivery system. But in reality this is not always the case.

With the procedures proposed in this draft the packet(s) received by the FHR will be dropped until the LHR has learned about the source and the SPT tree is built. That means for bursty sources or applications sensitive for the delivery of the first packet this proposal would not be very applicable. This proposal is mostly useful for applications that don't have strong dependency on the initial packet(s) and have a fairly constant data rate, like video distribution for example. For applications with strong dependency on the initial packet(s) we recommend using PIM Bidir [RFC5015] or SSM [RFC4607]. The protocol operations are much simpler compared to PIM SM, it will cause less churn in the network and both guarantee best effort delivery for the initial packet(s).

Another solution to address the problems described above is documented in [I-D.ietf-magma-msnip]. This proposal allows for a host to tell the FHR its willingness to act as Source for a certain Group before sending the data packets. LHRs have time to join the

SPT tree before the host starts sending which would avoid packet loss. The SG mappings announced by [I-D.ietf-magma-msnip] can be advertised directly in SG messages, allowing a very nice integration of both proposals. The life time of the SPT is not driven by the liveliness of Multicast data packets (which is the case with PIM SM), but by the announcements driven via [I-D.ietf-magma-msnip]. This will also prevent packet loss due to bursty sources.

7. Resiliency to network partitioning

In a PIM SM deployment where the network becomes partitioned, due to link or node failure, it is possible that the RP becomes unreachable to a certain part of the network. New sources that become active in that partition will not be able to register to the RP and receivers within that partition are not able to receive the traffic. Ideally you would want to have a candidate RP in each partition, but you never know in advance which routers will form a partitioned network. In order to be fully resilient, each router in the network may end up being a candidate RP. This would increase the operational complexity of the network.

The solution described in this document does not suffer from that problem. If a network becomes partitioned and new sources become active, the receivers in that partition will receive the SG Mappings and join the source tree. Each partition works independently of the other partition(s) and will continue to have access to sources within that partition. As soon as the network heals, the SG Mappings are re-flooded into the other partition(s) and other receives can join to the newly learned sources.

8. Security Considerations

The security considerations are no different from what is documented in [RFC5059].

9. IANA considerations

This document requires the assignment of a new PIM Protocol type for the PIM Flooding Protocol (PFP). IANA also needs to create a registry for PFP Types with type 0 allocated to "Source-Group Message". IANA also needs to create a registry for PFP TLVs, with type 0 allocated to the "Source Group Holdtime" TLV. The allocation procedures are yet to be determined.

10. Acknowledgments

The authors would like to thank Arjen Boers for contributing to the initial idea and Yiqun Cai for his comments on the draft.

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC5059] Bhaskar, N., Gall, A., Lingard, J., and S. Venaas, "Bootstrap Router (BSR) Mechanism for Protocol Independent Multicast (PIM)", RFC 5059, January 2008.

11.2. Informative References

- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.
- [I-D.ietf-magma-msnip] Fenner, B., Haberman, B., Holbrook, H., Kouvelas, I., and S. Venaas, "Multicast Source Notification of Interest Protocol (MSNIP)", draft-ietf-magma-msnip-06 (work in progress), March 2011.

Authors' Addresses

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

Email: ice@cisco.com

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

Email: stig@cisco.com

Michael Brig
Aegis BMD Program Office
17211 Avenue D, Suite 160
Dahlgren VA 22448-5148
USA

Email: michael.brig@mda.mil

PIM Working Group
Internet Draft
Intended status: Standards Track
Expires: January 13, 2014

Hong-Ke Zhang
Shuai Gao
Beijing Jiaotong University
T C.Schmidt
HAW Hamburg
Bo-hao Feng
Li-Li Wang
Beijing Jiaotong University
July 14, 2013

Multi-Upstream Interfaces IGMP/MLD Proxy
draft-zhang-pim-muiimp-01.txt

Abstract

In this document, followed by the idea mentioned in [1] and subsequently updated in [2], an IGMP/MLD proxy with multiple upstream interfaces called MUIIMP is proposed and analyzed. The MUIIMP inherits the basic rule of the IGMP/MLD proxy but extends with multiple upstream interfaces. To avoid data redundancy, each upstream interface of an MUIIMP device MUST NOT send or subscribe to the same data simultaneously.

Requirements Language

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

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January, 2014.

Copyright Notice

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

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

Table of Contents

1. Introduction.....	3
2. Terminology.....	3
3. MUIIMP Operation.....	4
3.1. The selection of default upstream interface.....	5
3.2. Report of downstream subscriptions on upstream interfaces..	5
3.3. Handover of the upstream interface.....	6
4. Use Case in PMIPv6 Environment.....	6
5. Security Considerations.....	9
6. References.....	9
Authors' Addresses.....	11
Acknowledgment.....	11

1. Introduction

RFC 4605 [3] specifies an IGMP/MLD proxy mechanism for forwarding based solely upon IGMP/MLD membership information in scenarios where multicast routing is not available. According to RFC 4605, an IGMP/MLD proxy performs the router portion of the IGMP/MLD protocol [4-5] on its downstream interfaces, and the host portion of the IGMP/MLD protocol on its single upstream interface.

The IGMP/MLD proxy mechanism can effectively extend the multicast scope and greatly simplify the implementation of edge devices. However, the IGMP/MLD proxy may exhibit inefficiency in some specific scenarios due to the limitation of single upstream interface. For example, in PMIPv6 multicast environment, multiple IGMP/MLD proxy instances need to be deployed at the MAG in RFC 6224 [6], which may result in tunnel convergence problem. In addition, there are also requirements to extend the IGMP/MLD proxy to support multiple upstream interfaces with the emergence of multi-homing.

It is noted that the idea about multiple upstream interfaces for IGMP/MLD proxy was firstly proposed in the draft [1] to improve the performance of multicast source mobility. Subsequently, the Multimob working group draft [2] describes the multi-upstream interfaces IGMP/MLD proxy in detail. Considering the multiple upstream interfaces extension is not only required for mobile multicast sources scenarios, this document is presented here.

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

2. Terminology

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

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

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

3. MUIIMP Operation

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

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

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

According to the role of the downstream nodes, the operation of the MUIIMP device will be described as follows:

1) Multicast listener on the downstream interface

Multicast listener reports are group-wise aggregated by the IGMP/MLD proxy. The aggregated report is issued to the upstream interface based on the subscriptions as well as the policy list. When receiving the IGMP/MLD subscriptions on the downstream interface, the MUIIMP decides whether to send the IGMP/MLD membership reports on the corresponding default upstream interface based on the membership database. The detailed membership subscriptions lookup and report decisions are discussed in Section 3.2.

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

2) Multicast source on the downstream interface

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

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

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

3.1. The selection of default upstream interface

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

The expression of the policy list is like below:

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

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

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

3.2. Report of downstream subscriptions on upstream interfaces

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

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

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

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

3.3. Handover of the upstream interface

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

4. Use Case in PMIPv6 Environment

With the development of the Mobile IP-like protocols (such as, MIPv6, PMIPv6 and DMM), combining multicast with mobile protocols has been widely discussed. One way is to deploy traditional IGMP/MLD proxy at the gateway of mobile nodes [2,6], but there are some drawbacks like the detour routing and tunnel convergence problem. MUIIMP can be used to optimize the behavior of multicast mobility as well as to support the multi-homing/multi-interface for both the mobile nodes and the fixed nodes. Requirements for an IGMP/MLD proxy with multiple interfaces have been proposed in the draft [8], covering a variety of application scenarios. In this section, a use-case in PMIPv6 environment is presented to illustrate how the MUIIMP works.

The basic deployment of MUIIMP in PIMv6 networks is shown in Figure 1, in which there are three mobile nodes (MNs) belonging to different LMAs and attaching under the same MAG. Let's adopt the scheme in RFC 6224 [6] and assume that the default upstream interface for each MN in Figure 1 is the tunnel interface from the MAG to the corresponding LMA. In real environment, the MN may also express its preference on upstream interface selection by other means. Detailed multicast related information of each MN is depicted

in Table 1. We assume that MN1, MN2 and MN3 attach at the MAG in sequence.

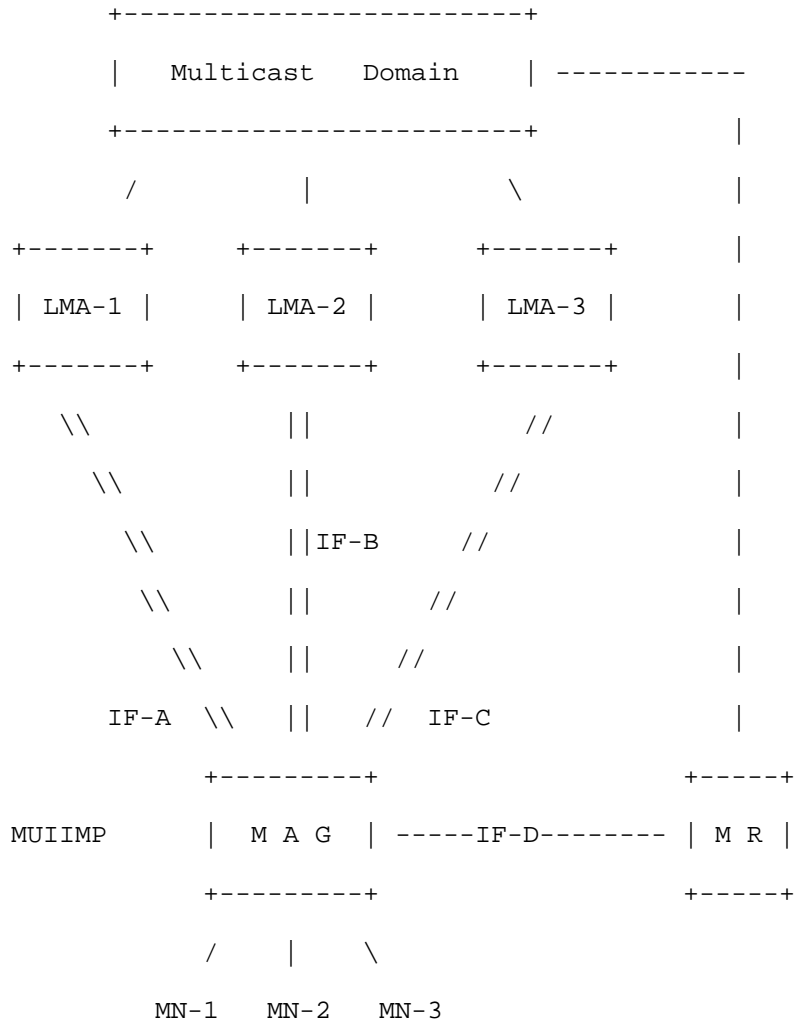


Figure 1: The basic deployment of MUIIMP in PIMv6 networks

Node	LMA	Type	Multicast Channel	Default U-IF
			(m1, EXCLUDE, { })	

MN-1	LMA-1	Receiver	(m2, EXCLUDE, { })	IF-A

			(m3, INCLUDE, {a})	
			(m1, EXCLUDE, { })	

MN-2	LMA-2	Receiver	(m2, INCLUDE, {b})	IF-B

			(m3, EXCLUDE, { })	
		Receiver	(m1, EXCLUDE, { })	
MN-3	LMA-3	-----	-----	IF-C
		Source	m2	

Table 1: Detailed information of each MN

The operation of the MUIIMP is described as follows:

- 1) MN1 firstly subscribes (m1, EXCLUDE, { }), (m2, EXCLUDE, { }) and (m3, INCLUDE, {a}), the MUIIMP initiates the corresponding IGMP/MLD subscription through the IF-A.

- 2) When MN2 (as well as MN3) subscribes (m1,EXCLUDE,{}), since there exists the same IGMP/MLD subscription through the IF-A, the MUIIMP will not initiate extra IGMP/MLD subscription.
- 3) When MN2 subscribes (m2,INCLUDE,{b}), since (m2,EXCLUDE,{}) has been subscribed, the MUIIMP does not initiate extra IGMP/MLD subscription.
- 4) When MN2 subscribes (m3,EXCLUDE,{}), since (m3,INCLUDE,{a}) has been subscribed through the IF-A, the MUIIMP will send IGMP/MLD subscription of (m3,EXCLUDE,{a}) through the IF-B.
- 5) When MN3 acts one of the sources for m2, traffic from MN3 are transmitted to the downstream interface towards MN1 as well as the IF-C. Besides, IGMP/MLD subscription of (m2,EXCLUDE,{}) sent by the MUIIMP through the IF-A is modified as (m2,EXCLUDE,{MN3}).
- 6) When MN1 hands over, the tunnel interface IF-A will be deleted and the MUIIMP needs to deal with this scenario because the subscriptions of MN2 and MN3 was sent through the IF-A previously. Here, IF-B and IF-C, as the default upstream interfaces of MN2 and MN3, are two candidate upstream interfaces for m1 subscription. Then, the predefined customized policy can be used to select a new U-IF from the candidate upstream interfaces. As for (m2,INCLUDE,{b}) and (m3,EXCLUDE,{}), they are only subscribed by the MN2, thus, the IF-B will be selected as the new U-IF for them.

It is worth noting that the fixed node (FN) connecting to the MAG can be regarded as an MN that never handovers. The default upstream interface of FN can be an interface adjacent to multicast domain, like the IF-D in Figure 1.

5. Security Considerations

To be done.

6. References

- [1] H. Zhang, Z. Yan, S. Gao, L. Wang, Q. Wu and H. Li, "Multicast Source Mobility Support in PMIPv6 Network", draft-zhang-multimob-msm-03, July 2011.
- [2] T C. Schmidt, S. Gao, H. Zhang and M. Waehlich, "Mobile Multicast Sender Support in Proxy Mobile IPv6 (PMIPv6) Domains", draft-ietf-multimob-pmipv6-source-03, February 2013.

- [3] B. Fenner, H. He, B. Haberman and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.
- [4] Cain, B., Deering, S., Kouvelas, I., Fenner, B. and A. Thyagarajan, "Internet Group Management Protocol, Version3", RFC 3376, October 2002.
- [5] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [6] T. Schmidt, M. Waehlis and S. Krishnan, "Base Deployment for Multicast Listener Support in Proxy Mobile IPv6 (PMIPv6) Domains", RFC 6224, April 2011.
- [7] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2000.
- [8] LM. Contreras, CJ. Bernardos and JC. Zuniga, "Extension of the MLD proxy functionality to support multiple upstream interfaces", draft-contreras-multimob-multiple-upstreams-01, February 2013.

Authors' Addresses

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

Phone: +861051684274
Email: hkzhang@bjtu.edu.cn
shgao@bjtu.edu.cn
11111021@bjtu.edu.cn
liliwang@bjtu.edu.cn

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

Email: schmidt@informatik.haw-hamburg.de
URI: <http://inet.cpt.haw-hamburg.de/members/schmidt>

Acknowledgment

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

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

W. Zhou
cisco Systems
February 22, 2013

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

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on August 26, 2013.

Copyright Notice

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

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

Table of Contents

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

1. Introduction

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

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

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

2. Tracking and Failover

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

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

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

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

3. PIM Assert Metric Auto-Adjustment

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

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

4. DF Election for BiDir Group

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

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

5. Tracking Multiple VRRP Groups on an Interface

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

6. Support of HSRP

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

7. Security Considerations

The proposed tracking mechanism has no negative impact on security.

8. Acknowledgments

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

9. Informative References

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

Author's Address

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

Email: weizho2@cisco.com

