PIM WG
Internet-Draft
Intended status: Standards Track
Expires: August 21, 2021

Z. Zhang
ZTE Corporation
F. Hu
Individual
B. Xu
ZTE Corporation
M. Mishra
Cisco Systems

February 17, 2021

Protocol Independent Multicast - Sparse Mode (PIM-SM) Designated Router (DR) Improvement
draft-ietf-pim-dr-improvement-11

Abstract

Protocol Independent Multicast - Sparse Mode (PIM-SM) is a widely deployed multicast protocol. As deployment for the PIM protocol is growing day by day, a user expects lower packet loss and faster convergence regardless of the cause of the network failure. This document defines an extension to the existing protocol, which improves the PIM's stability with respect to packet loss and convergence time when the PIM Designated Router (DR) role changes.

Table of Contents

1. Introduction 2
1.1. Keywords 3
2. Terminology 3
3. Protocol Specification 4
3.1. Election Algorithm 5
3.2. Sending Hello Messages 7
3.3. Receiving Hello Messages 8
3.4. Working with the DRLB function 8
4. PIM Hello message format 8
4.1. DR Address Option format 9
4.2. BDR Address Option format 9
4.3. Error handling 10
5. Backwards Compatibility 10
6. Security Considerations 10
7. IANA Considerations 11
8. Acknowledgements 11
9. References 11
9.1. Normative References 11
9.2. Informative References 12
Authors' Addresses 12

1. Introduction

Multicast technology, with PIM-SM ([RFC7761]), is used widely in Modern services. Some events, such as changes in unicast routes, or a change in the PIM-SM DR, may cause the loss of multicast packets.

The PIM DR has two responsibilities in the PIM-SM protocol. For any active sources on a LAN, the PIM DR is responsible for registering with the Rendezvous Point (RP). Also, the PIM DR is responsible for tracking local multicast listeners and forwarding data to these listeners.

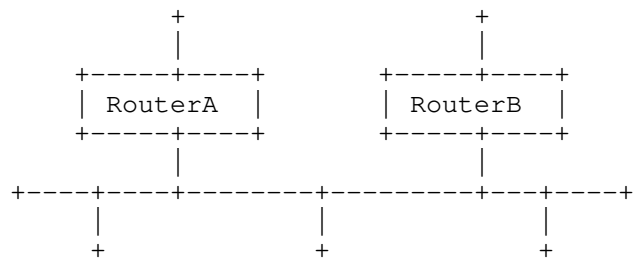


Figure 1: An example of multicast network

The simple network in Figure 1 presents two routers (A and B) connected to a shared-media LAN segment. Two different scenarios are described to illustrate potential issues.

(a) Both routers are on the network, and RouterB is elected as the DR. If RouterB then fails, multicast packets are discarded until RouterA is elected as DR, and it assumes the multicast flows on the LAN. As detailed in [RFC7761], a DR's election is triggered after the current DR's Hello_Holdtime expires. The failure detection and election procedures may take several seconds. That is too long for modern multicast services.

(b) Only RouterA is initially on the network, making it the DR. If RouterB joins the network with a higher DR Priority. Then it will be elected as DR. RouterA will stop forwarding multicast packets, and the flows will not recover until RouterB assumes them.

In either of the situations listed, many multicast packets may be lost, and the quality of the services noticeably affected. To increase the stability of the network this document introduces the Designated DR (DR) and Backup Designated Router (BDR) options, and specifies how the identity of these nodes is explicitly advertised.

1.1. Keywords

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Terminology

Modern services: The real time multicast services, such like IPTV, Net-meeting, etc.

Backup Designated Router (BDR): Immediately takes over all DR functions ([RFC7761]) on an interface once the DR is no longer present. A single BDR SHOULD be elected per interface.

Designated Router Other (DROther): A router which is neither a DR nor a BDR.

0x0: 0.0.0.0 if IPv4 addresses are in use or 0:0:0:0:0:0:0:0/128 if IPv6 addresses are in use. To simplify, 0x0 is used in abbreviation in this draft.

Sticky: The DR doesn't change unnecessarily when routers, even with higher priority, go down or come up.

3. Protocol Specification

The router follows the following procedures, these steps are to be used when a router starts, or the interface is enabled:

(a). When a router first starts or its interface is enabled, it includes the DR and BDR Address options with the OptionValue set to 0x0 in its Hello messages (Section 4). At this point the router considers itself a DROther, and starts a timer set to Default_Hello_Holdtime [RFC7761].

(b). When the router receives Hello messages from other routers on the same shared-media LAN, the router checks the value of DR/BDR address option. If the value is filled with a non-zero IP address, the router stores the IP address.

(c). After the timer expires, the router first executes the algorithm defined in section 3.1. After that, the router acts as one of the roles in the LAN: DR, BDR, or DROther.

If the router is elected the BDR, it takes on all the functions of a DR as specified in [RFC7761], but it SHOULD NOT actively forward multicast flows or send a register message to avoid duplication.

If the DR becomes unreachable on the LAN, the BDR MUST take over all the DR functions, including multicast flow forwarding and sending the Register messages. Mechanisms outside the scope of this specification, such as [I-D.ietf-pim-bfd-p2mp-use-case] or BFD Asynchronous mode [RFC5880] can be used for faster failure detection.

For example, there are three routers: A, B, and C. If all three were in the LAN, then their DR preference would be A, B, and C, in that order. Initially, only C is on the LAN, so C is DR. Later, B joins;

C is still the DR, and B is the BDR. Later A joins, then A becomes the BDR, and B is simply DROther.

3.1. Election Algorithm

The DR and BDR election refers the DR election algorithm defined in section 9.4 in [RFC2328], and updates the election function defined in section 4.3.2 in [RFC7761].

- The DR is elected among the DR candidates directly. If there is no DR candidates, i.e., all the routers advertise the DR Address options with zero OptionValue, the elected BDR will be the DR. And then the BDR is elected again from the other routers in the LAN.

- The BDR election is not sticky. Whatever there is a router that advertise the BDR Address option, the router which has the highest priority, expect for the elected DR, is elected as the BDR. That is the BDR may be the router which has the highest priority in the LAN.

- The advertisement is through PIM Hello message.

Except for the information recorded in section 4.3.2 in [RFC7761], the DR/BDR OptionValue from the neighbor is also recorded:

- neighbor.dr: The DR Address OptionValue that presents in the Hello message from the PIM neighbor.

- neighbor.bdr: The BDR Address OptionValue that presents in the Hello message from the PIM neighbor.

The pseudocode is shown below: A BDR election function is added, and the DR function is updated. The validneighbor function means that a valid Hello message has been received from this neighbor.

```
BDR(I) {
    bdr = NULL
    for each neighbor on interface I {
        if ( neighbor.bdr != NULL ) {
            if (validneighbor (neighbor.bdr) == TRUE) {
                if bdr == NULL
                    bdr = neighbor.bdr
                else (dr_is_better( neighbor.bdr, bdr, I ) == TRUE ) {
                    bdr = neighbor.bdr
                }
            }
        }
    }
    return bdr
}

DR(I) {
    dr = NULL
    for each neighbor on interface I {
        if ( neighbor.dr != NULL ) {
            if (validneighbor (neighbor.dr) == TRUE) {
                if (dr == NULL)
                    dr = neighbor.dr
                else (dr_is_better( neighbor.dr, dr, I ) == TRUE ) {
                    dr = neighbor.dr
                }
            }
        }
    }
    if (dr == NULL) {
        dr = bdr
    }
    if (dr == NULL) {
        dr = me
    }
    return dr
}
```

Compare to the DR election function defined in section 4.3.2 in [RFC7761] the differences include:

- The router, that can be elected as DR, has the highest priority among the DR candidates. The elected DR may not be the one that has the highest priority in the LAN.

- The router that supports the election algorithm defined in section 3.1 MUST advertise the DR Address option defined in section 4.1 in PIM Hello message, and SHOULD advertise the BDR Address option

defined in section 4.2 in PIM Hello message. In case a DR is elected and no BDR is elected, only the DR Address option is advertised in the LAN.

3.2. Sending Hello Messages

When PIM is enabled on an interface or a router first starts, Hello messages MUST be sent with the OptionValue of the DR Address option set to 0x0. The BDR Address option SHOULD also be sent, the OptionValue MUST be set to 0x0. Then the interface starts a timer which value is set to Default_Hello_Holdtime. When the timer expires, the DR and BDR will be elected on the interface according to the DR election algorithm (Section 3.1).

After the election, if there is one existed DR in the LAN, the DR remains unchanged. If there is no existed DR in the LAN, a new DR is elected, the routers in the LAN MUST send the Hello message with the OptionValue of DR Address option set to the elected DR. If there are more than one routers with non-zero DR priority in the LAN, a BDR is also elected. Then the routers in the LAN MUST send the Hello message with the OptionValue of BDR Address option set to the elected BDR. Any DROther router MUST NOT use its IP addresses in the DR/BDR Address option.

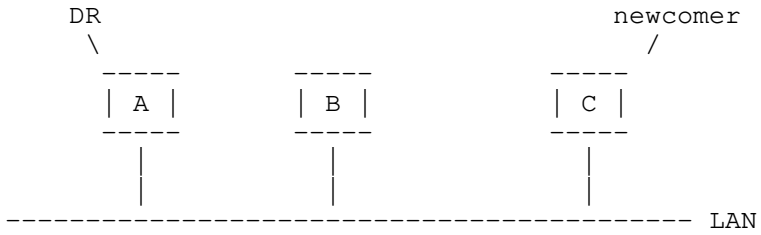


Figure 2

For example, there is a stable LAN that includes RouterA and RouterB. RouterA is the DR that has the highest priority. RouterC is a newcomer. RouterC sends a Hello message with the OptionValue of DR/ BDR Address option set to zero. RouterA and RouterB sends the Hello message with the DR OptionValue set to RouterA, the BDR OptionValue set to RouterB.

In case RouterC has a higher priority than RouterB, RouterC elects itself as the BDR after it runs the election algorithm, then RouterC sends Hello messages with the DR OptionValue set to the IP address of current DR (RouterA), and the BDR OptionValue set to RouterC.

In case RouterB has a higher priority than RouterC, RouterC finds that it can not be the BDR after it runs the election algorithm, it

sets the status to DROther. Then RouterC sends Hello messages with the DR OptionValue set to RouterA and the BDR OptionValue set to RouterB.

3.3. Receiving Hello Messages

When a Hello message is received, the OptionValue of DR/BDR is checked. If the OptionValue of DR is not zero and it isn't the same with local stored values, or the OptionValue of DR is zero but the advertising router is the stored DR, the interface timer of election MAY be set/reset.

Before the election algorithm runs, the validity check MUST be done. The DR/BDR OptionValue in the Hello message MUST match with a known neighbor, otherwise the DR/BDR OptionValue can not become the DR/BDR candidates.

If there is one or more candidates which are different from the stored DR/BDR value after the validity check, the election MUST be taken. The new DR/BDR will be elected according to the rules defined in section 3.1.

3.4. Working with the DRLB function

A network can use the enhancement described in this document with the DR Load Balancing (DRLB) mechanism [RFC8775]. The DR MUST send the DRLB-List Hello Option defined in [RFC8775]. If the DR becomes unreachable, the BDR will take over all the multicast flows on the link, which may result in duplicated traffic as it may not have been a Group DR (GDR). The new DR MUST then follow the procedures in [RFC8775].

In case the DR, or the BDR which becomes DR after the DR failure, doesn't support the mechanism defined in [RFC8775], the DRLB-List Hello Option can not be advertised, then the DRLB mechanism takes no effection.

4. PIM Hello message format

Two new PIM Hello Options are defined, which conform to the format defined in [RFC7761].

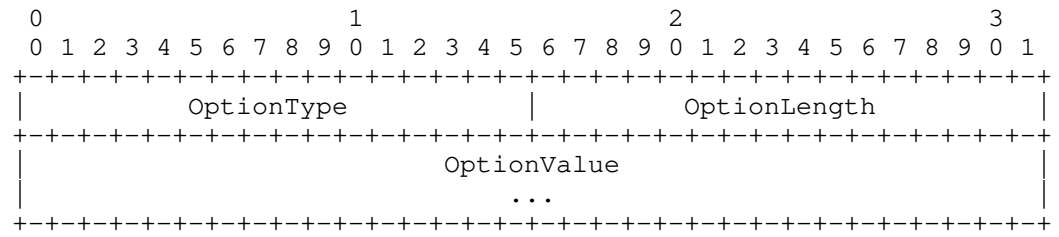


Figure 3: Hello Option Format

4.1. DR Address Option format

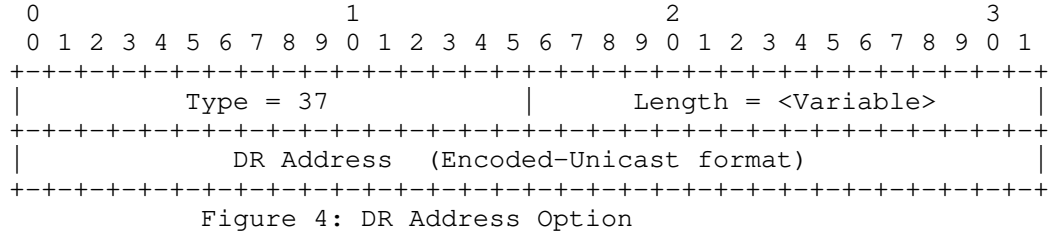


Figure 4: DR Address Option

- OptionType : The value is 37.
- OptionLength: 4 bytes if using IPv4 and 16 bytes if using IPv6.
- DR Address: If the IP version of the PIM message is IPv4, the value MUST be the IPv4 address of the DR. If the IP version of the PIM message is IPv6, the value MUST be the link-local address of the DR.

4.2. BDR Address Option format

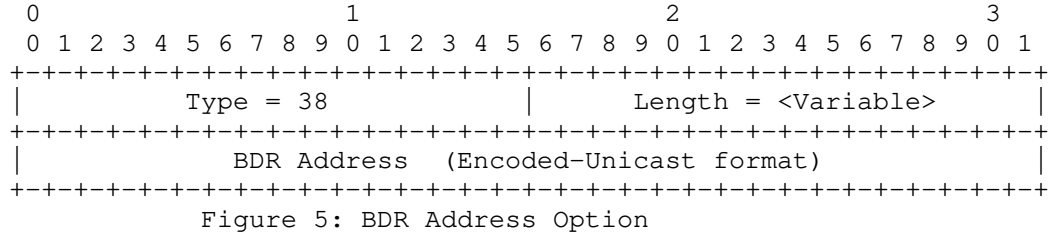


Figure 5: BDR Address Option

- OptionType : The value is 38.
- OptionLength: 4 bytes if using IPv4 and 16 bytes if using IPv6.
- BDR Address: If the IP version of the PIM message is IPv4, the value MUST be the IPv4 address of the BDR. If the IP version of

the PIM message is IPv6, the value MUST be the link-local address of the BDR.

4.3. Error handling

The DR and BDR addresses MUST correspond to an address used to send PIM Hello messages by one of the PIM neighbors on the interface. If that is not the case then the OptionValue of DR/BDR MUST be ignored as described in section 3.3.

An option with unexpected values MUST be ignored. For example, a DR Address option with an IPv4 address received while the interface only supports IPv6 is ignored.

5. Backwards Compatibility

Any router using the DR and BDR Address Options MUST set the corresponding OptionValues. If at least one router on a LAN doesn't send a Hello message, including the DR Address Option, then the specification in this document MUST NOT be used. For example, the routers in a LAN all support the options defined in this document, the DR/BDR is elected. A new router which doesn't support the options joins, when the hello message without DR Address Option is received, all the router MUST switch the election function back immediately. This action results in all routers using the DR election function defined in [RFC7761] or [I-D.mankamana-pim-bdr]. Both this draft and the draft [I-D.mankamana-pim-bdr], introduce a backup DR. The later draft does this without introducing new options but does not consider the sticky behavior. In case there is router which doesn't support the DR/BDR Address Option defined in this document, the routers SHOULD take the function defined in [I-D.mankamana-pim-bdr] if all the routers support it, otherwise the router SHOULD used the function defined in [RFC7761].

A router that does not support this specification ignores unknown options according to section 4.9.2 defined in [RFC7761]. So the new extension defined in this draft will not influence the stability of neighbors.

6. Security Considerations

[RFC7761] describes the security concerns related to PIM-SM. A rogue router can become the DR/BDR by appropriately crafting the Address options to include a more desirable IP address or priority. Because the election algorithm makes the DR role be non-preemptive, an attacker can then take control for long periods of time. The effect of these actions can result in multicast flows not being forwarded (already considered in [RFC7761]).

Some security measures, such as IP address filtering for the election, may be taken to avoid these situations. For example, the Hello message received from an untrusted neighbor is ignored by the election process.

7. IANA Considerations

IANA is requested to allocate two new code points from the "PIM-Hello Options" registry.

Type	Description	Reference
37	DR Address Option	This Document
38	BDR Address Option	This Document

Table 1

8. Acknowledgements

The authors would like to thank Alvaro Retana, Greg Mirsky, Jake Holland, Stig Venaas for their valuable comments and suggestions.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8775] Cai, Y., Ou, H., Vallepalli, S., Mishra, M., Venaas, S., and A. Green, "PIM Designated Router Load Balancing", RFC 8775, DOI 10.17487/RFC8775, April 2020, <https://www.rfc-editor.org/info/rfc8775>.

9.2. Informative References

[I-D.ietf-pim-bfd-p2mp-use-case]
Mirsky, G. and J. Xiaoli, "Bidirectional Forwarding Detection (BFD) for Multi-point Networks and Protocol Independent Multicast - Sparse Mode (PIM-SM) Use Case", draft-ietf-pim-bfd-p2mp-use-case-05 (work in progress), November 2020.

[I-D.mankamana-pim-bdr]
mishra, m., Goh, J., and G. Mishra, "PIM Backup Designated Router Procedure", draft-mankamana-pim-bdr-04 (work in progress), April 2020.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <https://www.rfc-editor.org/info/rfc5880>.

Authors' Addresses

Zheng(Sandy) Zhang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: zhang.zheng@zte.com.cn

Fangwei Hu
Individual
Shanghai
China

Email: hufwei@gmail.com

Benchong Xu
ZTE Corporation
No. 68 Zijinghua Road, Yuhuatai Distinct
Nanjing
China

Email: xu.benchong@zte.com.cn

Mankamana Mishra
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: mankamis@cisco.com

PIM Working Group
Internet-Draft
Intended status: Informational
Expires: May 13, 2019

LM. Contreras
Telefonica
CJ. Bernardos
Universidad Carlos III de Madrid
H. Asaeda
NICT
N. Leymann
Deutsche Telekom

November 9, 2018

Requirements for the extension of the IGMP/MLD proxy functionality to support multiple upstream interfaces
draft-ietf-pim-multiple-upstreams-reqs-08

Abstract

The purpose of this document is to define the requirements for a MLD (for IPv6) or IGMP (for IPv4) proxy with multiple interfaces covering a variety of applicability scenarios. The referred scenarios, while describing not sophisticated service situations, present cases that existing technology does not allow to solve in a simplistic manner. This document is then intended to serve as input for future documents defining the support of multiple upstream interfaces by IGMP/MLD proxies being compliant with the aforementioned requirements.

Table of Contents

1. Introduction 2
2. Terminology 3
3. Problem statement 3
4. Scenarios of applicability 5
4.1. Fixed network scenarios 6
4.1.1. Multicast wholesale offer for residential services . 6
4.1.1.1. Requirements 6
4.1.2. Multicast resiliency 7
4.1.2.1. Requirements 7
4.1.3. Load balancing for multicast traffic in the metro segment 7
4.1.3.1. Requirements 8
4.1.4. Network merging with different multicast services . . 8
4.1.4.1. Requirements 8
4.1.5. Multicast service migration 9
4.1.5.1. Requirements 9
4.2. Mobile network scenarios 10
5. Summary of requirements 10
6. Security Considerations 11
7. IANA Considerations 11
8. Acknowledgements 12
9. References 12
9.1. Normative References 12
9.2. Informative References 12
Authors' Addresses 14

1. Introduction

The aim of this document is to define the functionality that an IGMP/MLD proxy with multiple upstream interfaces should have in order to support different scenarios of applicability in both fixed and mobile networks. IGMP/MLD proxis are a generic solution very much deployed in existing carrier networks. An extension to them in the sense of supporting multiple upstream interfaces can provide a more flexible and lightweight solution than other potential alternatives that could face more complexities (like multi-domain routing in the case of PIM,

or the need of some external elements –e.g., controllers– if the coordination of actions required lays outside the proxy).

The functional behavior of an IGMP/MLD proxy with multiple upstream interfaces here described 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.

For doing that, a number of scenarios are described, representing current deployments and needs from operator's networks. From that scenarios, certain requirements are identified as needed to simplify operational situations, enable optimized service delivery, etc. Those represent functional requirements to be satisfied by IGMP/MLD proxies with multiple upstream interfaces. These functional requirements reflect the need of coordinating actions from a single element in the network (i.e., the IGMP/MLD proxy), optimizing the delivery of the content within the network at any time.

Any Source Multicast (ASM) [RFC1112] and Source-Specific Multicast (SSM) [RFC4607] represent different service models at the time of subscribing to multicast groups by means of IGMPv3 [RFC3376], [RFC5790] and MLDv2 [RFC3810]. When using ASM a receiver joins a group indicating only the desired group address to be received. In the case of SSM, a receiver indicates the specific source address as well as a group address from where the multicast content is received. Both service models are taken into account along this document, and the specific requirements are derived from them.

2. Terminology

This document uses the terminology defined in [RFC4605]. Specifically, the definition of Upstream and Downstream interfaces, which are repeated 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 IGMP/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.

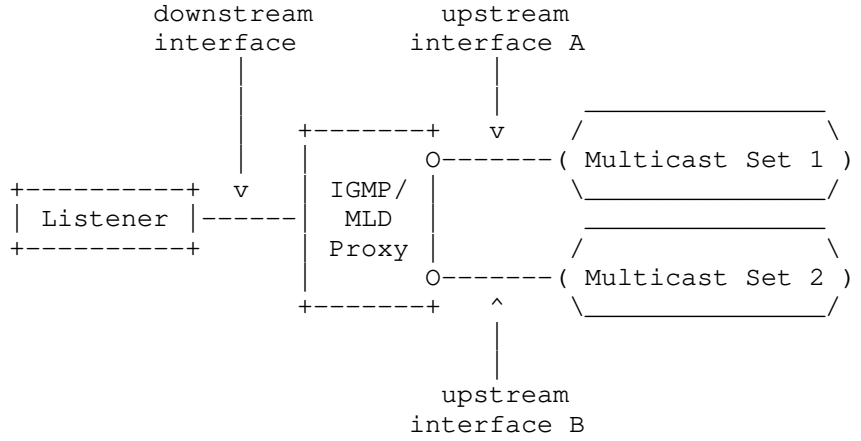


Figure 1: Concept of IGMP/MLD proxy with multiple upstream interfaces

This document is focused on both fixed and mobile network scenarios. Applicability of IGMP/MLD proxies with multiple upstream interfaces in mobile environments has been previously identified as beneficial in scenarios as the ones described in [RFC6224] and [RFC7287].

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

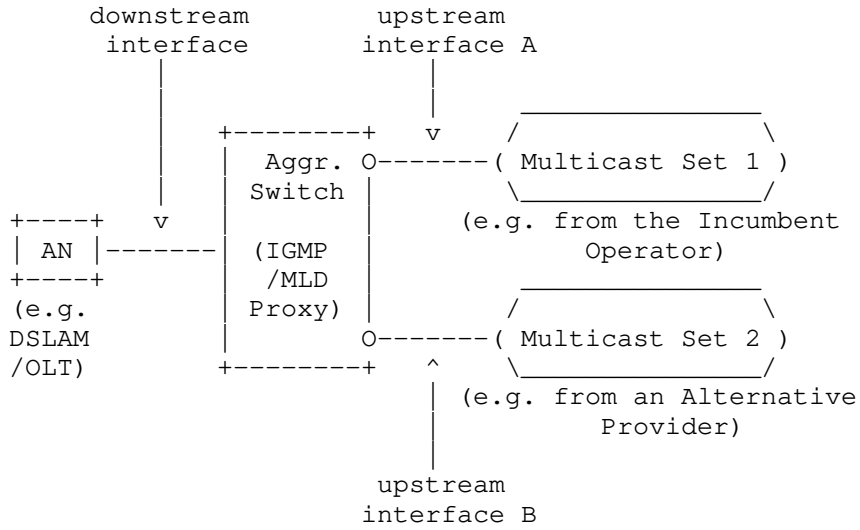


Figure 2: Example of usage of an IGMP/MLD proxy with multiple upstream interfaces in a fixed network scenario

Since those scenarios can motivate distinct needs in terms of IGMP/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 IGMP/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

Having multiple upstream interfaces creates a new decision space for delivering the proper multicast content to the subscriber. Basically it is now possible to implement channel-based (i.e., leveraging on multicast group IP address) or subscriber-based (i.e., referenced to the subscriber IP address) upstream selection, according to mechanisms or policies that could be defined for the multicast service provisioning.

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

All the exemplary scenarios here described are based on the support of two upstream interfaces. However, all of them are applicable also to the support of more than two upstream interfaces.

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 IGMP/MLD proxy with multiple upstream interfaces in those switches appears as a simple and straightforward solution.

4.1.1. Multicast wholesale offer for residential services

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

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

4.1.1.1. Requirements

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

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

- The IGMP/MLD proxy should be able to support ASM and SSM at the time of requesting the content. Since the use case assumes that each provider offers distinct multicast groups, the IGMP/MLD proxy should be able to identify inconsistencies in the SSM requests, that is, the case in which for an (S, G) request the source S does not deliver a the group G.

4.1.2. Multicast resiliency

In current PIM-based solutions [RFC7063], the resiliency of the multicast distribution relays on the routing capabilities provided by protocols like PIM [RFC7761] and VRRP [RFC5798]. A simpler scheme could be achieved by implementing different upstream interfaces on IGMP/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 mode for fast switching. The objective is to avoid video delivery affection that could imply play out interruption or buffering on the user side. Service parameters like the ones defined in [Y.1540] (such as packet loss ratio) or in [RFC4445] (like the delay factor) can be considered as parameters to be assesed from the service perspective. For instance, [TECH.3361-1] could be considered as a SLA framework to be satisfied in this case.

4.1.2.1. Requirements

- The IGMP/MLD proxy should be able to deliver multicast control messages received in the active upstream to the end users, while ignoring the control messages of the standby upstream interface.

- The IGMP/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.

- The IGMP/MLD proxy should be able to deliver IGMP/MLD messages sent by the end user (for both ASM and SSM modes) to the corresponding active upstream interface.

4.1.3. Load balancing for multicast traffic in the metro segment

A single upstream interface in existing IGMP/MLD proxy functionality [RFC4605] typically forces the distribution of all the channels on the same path in the last segment of the network. The metro and backhaul network is usually built using ring topologies. The devices in the ring implement IGMP/MLD functionality to join the content. Multiple upstream interfaces could naturally help to split the content demand, alleviating the bandwidth requirements in the overall metro segment by allowing some of the channels to follow the protection path, where spare capacity is vacant under normal conditions. This will allow, for instance, to absorb traffic peaks when a high number of channels (more than the expected on average) is requested.

4.1.3.1. Requirements

- The IGMP/MLD proxy should be able to deliver multicast control messages sent by the end user to the corresponding multicast router which provides the channel of interest.

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

- The IGMP/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).

- In the case of ASM, the IGMP/MLD proxy should be able to balance the traffic as a function of the group G requested. In the case of SSM, the load balancing mechanism could also consider the source S for the decision. In any case, the criteria will follow the olicies defined by the network operator. Such policies can be influenced by the user requesting the service, for instance through the subscription to some channels being offered by a third party (which has reached an agreement with the provider for delivering that content in its network).

4.1.4. Network merging with different multicast services

In some network merging situations, the multicast services provided before in each of the merged networks are maintained for the respective customer base (usually in a temporal fashion until the multicast service is redefined in a new single offer, but not necessarily, or not in short term, e.g. because of commercial agreements for each of the previous service offers).

In order to assist that network merging situations, IGMP/MLD proxies with multiple upstream interfaces can help in the transition simplifying the service provisioning and facilitating service continuity.

4.1.4.1. Requirements

- The IGMP/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, according to the service subscription.

- The IGMP/MLD proxy should be able to deliver multicast control messages sent by each of the multicast routers to the corresponding end user, according to the service subscription.

- The IGMP/MLD proxy should be able to decide which upstream interface is selected for any new channel request according to defined criteria (e.g., service subscription).

- For this use case, the usage of SSM can simplify the decision of the IGMP/MLD proxy. For ASM the decision should be assisted by further information like the service to which the end user is subscribed (e.g., taking into account what is the original network from where the end user was part previous to the network merge situation).

4.1.5. Multicast service migration

This scenario considers the situation where a multicast service needs to be migrated from one upstream interface to another upstream interface (e.g. because of changes inside the service provider's network). The migration should be "smooth" and without any service interruption. In this case the multicast content is initially offered in both upstream interfaces and the proxy dynamically switches from the first to the second upstream interface, according to certain policies, and enabling to shut down the first upstream interface once the migration is completed.

4.1.5.1. Requirements

- The IGMP/MLD proxy should be able to deliver multicast control messages sent by the end user to the corresponding multicast router before and after the service migration.

- The IGMP/MLD proxy should be able to deliver multicast control messages sent by each of the multicast routers to the corresponding end user, according to the situation of the user with respect to the service migration.

- The IGMP/MLD proxy should be able to decide which upstream interface corresponds to each user, according to the situation of the user with respect to the service migration, i.e., the status of the user with respect the platform migration as purely operational situation while transitioning from one platform to another in a smooth manner.

- The IGMP/MLD proxy should be able to decide which upstream interface corresponds to each ASM or SSM request, according to the situation of the group and source included in the request with respect to the service migration.

4.2. Mobile network scenarios

Mobile networks offer different alternatives for multicast distribution.

One of them is defined by 3GPP [TS23.246] for the Multimedia Broadcast Multicast Service (MBMS). In this case, a MBMS gateway (MBMS GW) is connected to multiple evolved Node B (eNodeB) -- which are the base stations connecting the mobile handsets with the network wirelessly [TS36.300] -- for data distribution by means of IP multicast. The MBMS GW delivers the IP multicast groups. The eNodeB joins the appropriate group multicast address allocated by the MBMS GW to receive the content data. At this distribution level, an IGMP/ MLD proxy could be part of the transport infrastructure providing connectivity to several distributed eNodeBs. The potential scenarios from this case do not essentially differentiate from the ones described for the fixed network scenarios, so the same situations and requirements apply.

Another alternative is given by Proxy Mobile IPv6 (PMIPv6) protocol for IP mobility management [RFC5213]. 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. As the basic solution for the distribution of multicast traffic within a PMIPv6 domain, [RFC6224] makes use of the bi-directional LMA-MAG tunnels. The use of an MLD proxy supporting multiple upstream interfaces can improve the performance and the scalability of multicast-capable PMIPv6 domains, for both multicast listener and multicast source mobility. Once again, the potential scenarios in this case are contained into the ones described for the fixed network scenarios, so the same situations and requirements apply.

5. Summary of requirements

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

Functio-nality	Multicast Wholesale	Multicast Resiliency	Load Balancing	Network Merging	Network Migration
Upstream Control Delivery	X	X	X	X	X
Downstr. Control Delivery	X	X	X	X	X
Active / Standby Upstream		X			
Upstr i/f selection per group			X	X	
Upstr i/f selection all group		X			X
ASM	X	X	X	X	X
SSM	X	X	X		X

Figure 3: Functionality needed on IGMP/MLD proxy with multiple upstream interfaces per application scenario

6. Security Considerations

All the security considerations in [RFC4605] are directly applicable to this proposal.

7. IANA Considerations

There are no IANA considerations.

8. Acknowledgements

The authors would like to thank (in alphabetical order) Alvaro Retana, Thomas C. Schmidt, Stig Venaas and Dirk von Hugo for their comments and suggestions.

9. References

9.1. Normative References

[RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, DOI 10.17487/RFC1112, August 1989, <https://www.rfc-editor.org/info/rfc1112>.

[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, DOI 10.17487/RFC4605, August 2006, <https://www.rfc-editor.org/info/rfc4605>.

[RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, DOI 10.17487/RFC4607, August 2006, <https://www.rfc-editor.org/info/rfc4607>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

9.2. Informative References

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <https://www.rfc-editor.org/info/rfc3376>.

[RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <https://www.rfc-editor.org/info/rfc3810>.

[RFC4445] Welch, J. and J. Clark, "A Proposed Media Delivery Index (MDI)", RFC 4445, DOI 10.17487/RFC4445, April 2006, <https://www.rfc-editor.org/info/rfc4445>.

[RFC5213] Gundavelli, S., Ed., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "Proxy Mobile IPv6", RFC 5213, DOI 10.17487/RFC5213, August 2008, <https://www.rfc-editor.org/info/rfc5213>.

[RFC5790] Liu, H., Cao, W., and H. Asaeda, "Lightweight Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Version 2 (MLDv2) Protocols", RFC 5790, DOI 10.17487/RFC5790, February 2010, <https://www.rfc-editor.org/info/rfc5790>.

[RFC5798] Nadas, S., Ed., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC 5798, DOI 10.17487/RFC5798, March 2010, <https://www.rfc-editor.org/info/rfc5798>.

[RFC6224] Schmidt, T., Waehlisch, M., and S. Krishnan, "Base Deployment for Multicast Listener Support in Proxy Mobile IPv6 (PMIPv6) Domains", RFC 6224, DOI 10.17487/RFC6224, April 2011, <https://www.rfc-editor.org/info/rfc6224>.

[RFC7063] Zheng, L., Zhang, J., and R. Parekh, "Survey Report on Protocol Independent Multicast - Sparse Mode (PIM-SM) Implementations and Deployments", RFC 7063, DOI 10.17487/RFC7063, December 2013, <https://www.rfc-editor.org/info/rfc7063>.

[RFC7287] Schmidt, T., Ed., Gao, S., Zhang, H., and M. Waehlisch, "Mobile Multicast Sender Support in Proxy Mobile IPv6 (PMIPv6) Domains", RFC 7287, DOI 10.17487/RFC7287, June 2014, <https://www.rfc-editor.org/info/rfc7287>.

[TECH.3361-1] European Broadcasting Union, "Service Level Agreement for media transport services", EBU TECH.3361-1, September 2014.

[TS23.246] "TS 23.246 Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (Release 14) V14.1.0.", 3GPP TS 23.246 V14.1.0 , December 2016.

[TS36.300] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2", 3GPP TS 36.300 10.11.0, September 2013.

[Y.1540] ITU-T, "Internet protocol data communication service - IP packet transfer and availability performance parameters", ITU-T Y.1540, July 2016.

Authors' Addresses

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

Email: luismiguel.contrerasmurillo@telefonica.com
URI: http://lmcontreras.com/

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/

Hitoshi Asaeda
National Institute of Information and Communications Technology
4-2-1 Nukui-Kitamachi
Koganei, Tokyo 184-8795
Japan

Email: asaeda@nict.go.jp

Nic Leymann
Deutsche Telekom
Germany

Email: n.leymann@telekom.de

BIER Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 9, 2017

IJ. Wijnands
P. Pfister
Cisco Systems
J. Zhang
Juniper Networks

July 8, 2016

Generic Multicast Router Election on LAN's
draft-wijnands-bier-mld-lan-election-01.txt

Abstract

When a host is connected to multiple multicast capable routers, each of these routers is a candidate to process the multicast flow for that LAN, but only one router should be elected to process it. This document proposes a generic multicast router election mechanism using Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) that can be used by any Multicast Overlay Signalling Protocol (MOSP). Having such generic election mechanism removes a dependency on Protocol Independent Multicast (PIM).

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 9, 2017.

Copyright Notice

Table of Contents

1. Introduction 2
2. Terminology and Definitions 3
3. Specification of Requirements 4
4. Problem Statement 4
4.1. Receiver side 4
4.2. Sender side 5
5. Proposal 6
6. DF Election Mechanism Requirements 7
7. The DF Election mechanism 8
7.1. Highest Random Weight 8
7.2. The DF Hello Message 8
7.3. The Designated Announcer 9
7.3.1. DAL Hello Option 9
7.3.2. A new Candidate DF 9
7.3.3. A candidate DF goes down 10
7.4. DA Inconsistency 11
8. The Hello Message Packet Format 11
9. Security Considerations 11
10. IANA Considerations 12
11. Acknowledgments 12
12. Normative References 12
Authors' Addresses 13

1. Introduction

Hosts connected to Local Area Networks (LAN) use Internet Group Management Protocol (IGMP) [RFC4605] or Multicast Listener Discovery (MLD) [RFC3810] to report their interest in a particular multicast flow. A multicast flow is identified by a Group or a combination of Group and Source address. Routers connected to a LAN listen to these membership reports and signal that information to the Multicast Overlay Signalling Protocol (MOSP). When a host is connected to multiple routers, each of these routers is a candidate to forward the multicast flow onto that LAN, but only one of them should forward the packets for a given flow to avoid duplication of Multicast packets. A similar requirement exists for hosts that are sending multicast traffic and are connected to multiple routers on a LAN. If multiple routers accept the multicast packets from the LAN, duplication may occur and/or routing loops may be created.

Protocol Independent Multicast (PIM) [RFC4601] is a MOSP and has a built-in mechanism to elect a Designated Router (DR) on the receiver LAN and a Designated Forwarder (DF) on the senders LAN. The DR/DF election avoids duplication and looping of multicast packets. Other existing or candidate MOSPs, like Border Gateway Protocol (BGP) [RFC6514], Multi-point Label Distribution Protocols (mLDP) [RFC6826], Locator ID Seperation Protocol (LISP) [RFC6830] and IGMP/MLD [I-D.pfister-bier-mld] have no embedded LAN DR/DF election mechanism. These MOSPs still rely on PIM to perform DR/DF election on LANs.

With the introduction of mLDP and Bit Indexed Explicit Replication (BIER) [I-D.ietf-bier-architecture], there is no dependency on PIM to transport multicast packets through the network. Having a dependency on PIM just for DR/DF election is undesirable if PIM is not selected as the MOSP. This document proposes a generic DR/DF election which can be used by any MOSP without having a dependency on PIM. It potentially allows for different MOSPs to coexistence on single LANs.

2. Terminology and Definitions

Readers of this document are assumed to be familiar with the terminology and concepts of the documents listed as Normative References. For convenience, some of the more frequently used terms appear below.

LAN:
: Local Area Network.

IGMP:
: Internet Group Management Protocol.

MLD:
: Multicast Listener Discovery.

mLDP:
: Multipoint LDP.

PIM:
: Protocol Independent Multicast.

ASM:
: Any Source Multicast.

RP:
: The PIM Rendezvous Point.

LISP:
: Locator ID Seperation Protocol.

BIER:
Bit Indexed Explicit Replication.

MOSP:
Multicast Overlay Signalling Protocol. This is a protocol that is (potentially) capable of announcing multicast flow membership across the network between multicast routers. For example PIM, mLDP, BGP, IGMP, MLD and LISP.

DF:
A Designated Forwarder is responsible for accepting a multicast packet from a LAN.

DR:
A Designated Router is responsible for forwarding a multicast packet onto a LAN.

DA:
A Designated Announcer is a router that is responsible for announcing a list of candidate Designated Forwarders.

DAL:
A Designated Announcer List is generated by the DA and holds the candidate Designated Forwarders.

3. Specification of Requirements

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

4. Problem Statement

In the following sections we describe the requirements for DR/DF election in more detail for hosts that are multicast senders and receivers connected to multiple routers on a single LAN.

4.1. Receiver side

Consider the network below in Topology1.

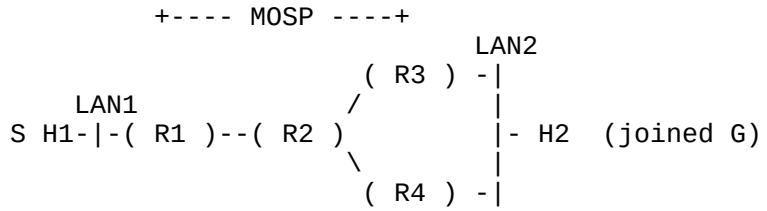


Figure 1

Suppose that H2 on LAN2 is joining a multicast Group G. The MOSP runs between R1, R3 and R4. Both R3 and R4 will receive the IGMP/MLD report, but only one of these should become the DR. One might consider that this problem can be detected and resolved by the MOSP. The MOSP could be enhanced to allow R1 to detect that both R2 and R4 are connected to the same LAN, and select only to forward the multicast flow to R3. That would solve the problem in the above topology, but would fail in the topology below:

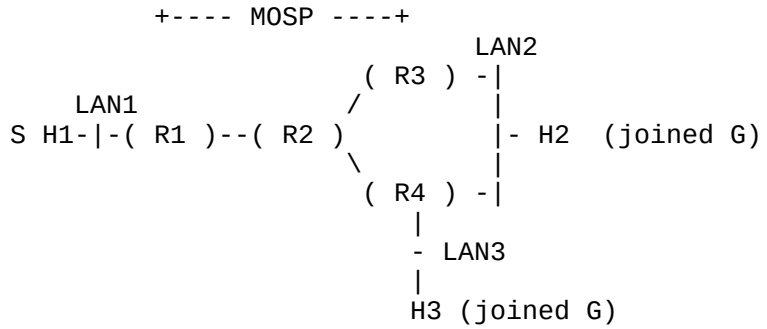


Figure 2

Consider that H3 on LAN3 joined the same multicast Group G. Since H3 is singly connected to R4, router R1 needs to forward the multicast flow to R4 in order for H3 to receive the packets. R4 does not have enough information to determine whether or not to forward on LAN2 for H2 when it receives the multicast packets due to H3. In other words, R4 needs DR state to avoid sending packets to H2 on LAN2.

4.2. Sender side

Consider the network below in Topology3.

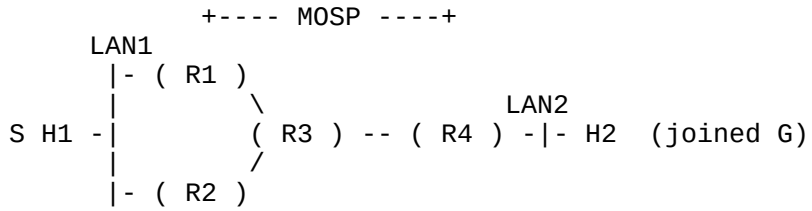


Figure 3

H1 is directly connected via a LAN1 to R1 and R2. H2 joins a multicast Group G, without specifying the Source. This is called Any Source Multicast (ASM). The MOSP signals R4s interest in Group G to R1 and R2. Note that there is no PIM deployed in this network and there is no Rendezvous Point (RP) that is a target for this receiving this Group membership. R4 has no information which routers in this network have multicast packets to sent for this Group. Since this is ASM, there may be multiple senders for this Group and H2 wants to receive them all. For that reason, R4 will use the MOSP to announce the membership to all edge nodes in the network (R1 and R2). This poses a potential problem since R1 and R2 are both directly connected to the Source S. If both R1 and R2 will forward the multicast packets to R4, H2 will receive duplicate packets. This is a problem that only occurs when a Source is dually connected to two or more routers connected to the sam LAN. This problem can be resolved by doing a Designated Forwarder (DF) election, similar to the DR election. If R1 and R2 are aware they are directly connected, an election will cause only one of them to forward the multicast packets into the network for a given (S,G) flow.

5. Proposal

As explain in Section 4, it is desirable to have a generic DR/DF election mechanism that can be used by existing and candidate MOSPs. Also, if a mix of MOSPs is used in the network, it is not obvious which MOSP is responsible for electing the DR/DF. If a single DR/DF is to be elected, and each MOSP does its own election, the MOSPs have to negotiate among each other who will be responsible for DR/DF on a LAN. Independent of the MOSP, a single router connected to the LAN should be elected. It seems inefficient and unpractical to have each MOSP implement its own DR/DF election mechanism.

There is a process in the router that all the MOSPs depend on for Group membership discovery, that is the IGMP/MLD process. The DR/DF election is typically based on the Group address or Group and Source address of the multicast flow. This information is available in the IGMP/MLD process. In this document we propose to enhance the IGMP/MLD protocol to allow a DR/DF election among multicast routers

connected to a LAN. As soon as a router is elected as DR/DF, it can select the MOSP that will be responsible to deliver the multicast flow to this router, and onwards onto the LAN(s).

IGMP/MLD has support for electing a Membership Querier based on the lowest IP address of the multicast routers sending out Membership Queries. It would be possible to use the elected Membership Querier as the DR/DF on a LAN. However, the authors believe that the Membership Querier procedures are not robust and extensible enough to be used DR/DF for election on LANs. For example, if a new multicast router becomes active on a LAN, it will immediately assume the role of a Membership Querier, which can lead to duplication and/or looping of packets if also used as DR/DF. This duplication/looping will last until it learns about other Membership queriers with a lower IP address. Having two Membership queriers on the LAN has limited impact on the IGMP/MLD protocol it self, it would only cause more Membership Reports to be received.

The election mechanism for the DR and the DF is very similar. In fact, when a DF is elected, it MUST always be used as the DR as well to avoid multicast packet looping. The procedures in this document always elect a DF on the LAN, and for that reason will always be the DR. In the sections that follow, we don’t refer to the DR anymore. Everywhere where we reference DF, we implicitly mean it applies to both the DR and DF.

6. DF Election Mechanism Requirements

When electing a DF on the LAN, it is important to have a single DF for a given Multicast flow at all times. If during the election process (or changes to it), there is no DF, it will cause traffic loss to the end user. If there are two (or more) DFs at the same time, it may cause traffic duplication or even loops. Since the election is done among different routers, it is not so trivial to guarantee that there will never be inconsistency in the DF election. There is also a tradeoff between the complexity introduced and the incremental benefit it brings. The procedures in this document are designed to detect inconsistency and recover from it as fast as possible. During inconsistency, we prefer traffic loss over possible duplication or looping of multicast packets.

When there are multiple candidate DF routers on the LAN, it is beneficial to load-balance the traffic over the different candidate DFs. This helps to distribute the bandwidth usage among the routers, reduce the impact of a router failure and shorten the failover time when changing the DF for effected flows. For that reason the DF procedures MUST support DF election per multicast group address.

7. The DF Election mechanism

7.1. Highest Random Weight

The method proposed to select a DF is based on the Highest Random Weight (HRM) as described in [RFC2291]. The paragraph below is mostly taken (and modified) from [RFC2291].

The router computes the weight for EACH candidate DF by performing a hash over the Group address that identifies the flow, as well as over the address of the candidate DF. The router then chooses the candidate DF with the highest resulting weight value. This has the advantage of minimizing the number of flows affected by a candidate DF addition or deletion (only 1/N of them), but is approximately N times as expensive as a modulo-N hash.

In order to get a good distribution of the Group addresses over the candidate DFs, it is important we choose a good Pseudorandom function to calculate the Weight. The Weight is calculated using the Group (G) IP address and the Candidate DF (CDF) IP address.

Weight(G, CDF) =
(1103515245((1103515245.G+12345)XOR CDF)+12345)(mod 2^31)

If multiple Candidate DFs end up with the same highest weight, the DF with the lowest IP address MUST be selected.

If every candidate DF on the LAN uses the same HRW algorithm to select the DF for a particular Group out of the same list of candidate DFs, they all will reach the same conclusion and there will be no inconsistency. It is very important every router on the LAN has the same list of candidate DFs. The mechanism proposed in this draft to generate a consistent list is based on the new Hello message.

7.2. The DF Hello Message

In order to discover the candidate DFs we need a mechanism to learn them. We introduce a new (IGMP/MLD) message type called the DF Hello. Routers on a LAN that are candidate DFs periodically send DF Hellos. The message format is specified later in a later revision document. Based on the DF Hellos it is possible to generate a list of candidate DFs. However, it is challenging to keep the candidate DF list synchronized between the routers when DFs are added or removed from the list as each router will do that based on its own scheduling. Especially when candidate DFs timeout, it is very likely this happens at different times and opens up the opportunity for inconsistency. Also, when a new candidate DF is added to the network

and one of the routers did not get the initial DF Hello message, its candidate DF list will be out of sync until the next DF Hello is received, leading to a inconsistent candidate DF list for a relatively long period. In order to help synchronize the candidate DF List we elect a Designated Announcer (DA).

7.3. The Designated Announcer

The router that will act as the Designated Announcer is determined by the Priority value as included in the Hello message, using the IP address as tiebreaker. The router with the highest priority is preferred, if there are multiple routers with the same priority, the router with the highest IP address is preferred. The DA determines which routers from the Hello List (HL) are included in the Designated Announcer List (DAL). By default all the routers in the HL are considered to be included in the DAL. It is however possible to filter certain candidates and not include these in the list based on some sort of preference.

7.3.1. DAL Hello Option

The DAL is sent out by the DA as an Option included in its Hello message. In order to reliably transmit the Hello Message with the DAL option, a DAL sequence number is included in the packet along with an acknowledgement flag for each router in the DAL. Every router in the DAL MUST respond by triggering a Hello message including this sequence number. If the DA has not received a response within a given timeout from certain routers in the DAL it will re-transmit the Hello message with the Acknowledgement flag not set for the routers that have not responded. The routers on the LAN that see their IP address in the DAL without the acknowledgement flag set will re-transmit their Hello. This process continues until the DA has received a response from all the routers in the DAL. Using this mechanism we minimize the time an inconsistency can occur when a router has missed a Hello message that includes that DAL.

7.3.2. A new Candidate DF

When a new candidate DF becomes active on the LAN, it first has to learn if there are other candidate DFs on the LAN. Learning about other candidate DFs is accelerated by setting the Learn Flag in the Hello message. Routers on the LAN that receive a Hello with the Learn Flag set will trigger a Hello message in response. After the learning delay the new DF assumes all candidate DFs on the LAN have responded and the Hello List is complete. There are three different scenarios the new DF has to consider.

7.3.2.1. The Hello List is empty

When the HL is empty, the new DF will become the DA with only its own address in the DAL. The DF will start to act as DF for all the groups.

7.3.2.2. The New DF is not the DA

When there are other candidate DFs on the LAN, the Hello List is populated. If the new DF is not the DA, it will have to wait for the DA to include its address in the DAL. As soon as it sees its own address in the ADA with the acknowledgement flag not set, it will trigger a Hello message with the DAL sequence number and start to act as DF. Note, it is likely that new DFs IP address is already included in the first Hello message it receives from the DA.

7.3.2.3. The New DF is the DA

After the Learning delay the new router may find it self having the highest Priority and will be the new ADA. Note, we prefer the DA to be deterministic so the new DF will take over the role of the DA. The DF which is currently the DA will have seen the Hello message from the new DF and will realize this is the new DA. The current DA MUST respond by sending a Hello message without the DAL in it. All the routers on the LAN will now know that the current DA is going away. The candidate DFs MAY continue to use the old DAL until the new DAL list is received from the new DA. The new DF will create the DAL list based on its Hello List and send out a Hello message, following the procedures as described above. If during a transition of the DA a router detects inconsistency between the received DAL and the perceived DA, the router stops using the current DAL and waits until the inconsistency is resolved. This inconsistency may have occurred due to missing a DF Hello message (also see section DA inconsistency).

7.3.3. A candidate DF goes down

When a DF goes down there are 2 different scenarios to consider.

7.3.3.1. The DF was the DA

When a DF goes down, due to a failure or an operator removing it from the LAN, the routers on the LAN will eventually detect this because the Holdtime for that DF will expire. This does not have an immediate effect on the DF procedures because the DF is chosen from the DAL, originated by the DA. A candidate DF MUST NOT take any action based on a candidate DF going down, but MUST wait for the DA to sent out a new DAL list. This will ensure that all candidate DFs

on the LAN will start to use the new DAL at the same time and avoid any discrepancies due to routers expiring the timer associated with the DF that went down.

7.3.3.2. The DF was the DA

If the DF that goes down is the DA, a new DA has to be elected. Note, every candidate DF on the LAN is a potential candidate to become the new DA. The new DA is chosen based on the Hello List using the Designated Announcer election procedures. It is possible a candidate DF receives the DAL from the new DA before it detected the current DA is down. This may be due to a race condition where timers on the candidate DF expire at different times. We use the procedures as described in section (DA inconsistency).

7.4. DA Inconsistency

A candidate DF that receives a DAL from a router that it does not consider to be the active DA MUST immediately stop acting as a DF. The candidate DF MUST wait for the DA inconsistency to be resolved before it is allowed to resume its role as candidate DF. This will cause traffic to be blocked for the multicast groups this DF is responsible for, but it will not cause traffic duplication and/or loops due to other DFs using a different DAL list. The inconsistency can be resolved due to the following events.

- The active DA expires.
- A Hello is received from the active DA without a DAL.

When the candidate DF detects that there is only one candidate DF that has announced the DAL and it is considered to be the DA, the inconsistency is resolved and the DF can resume its role as DF for the Groups it is responsible for.

8. The Hello Message Packet Format

The format of the Hello Message is included on the next revision of this document.

9. Security Considerations

TBD.

10. IANA Considerations

TBD.

11. Acknowledgments

Many thanks to Neale Ranns and Greg Shepherd for their comments on this draft.

12. Normative References

[I-D.ietf-bier-architecture]
Wijnands, I., Rosen, E., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast using Bit Index Explicit Replication", draft-ietf-bier-architecture-02 (work in progress), July 2015.

[I-D.pfister-bier-mld]
Pfister, P., Wijnands, I., and M. Stenberg, "BIER Ingress Multicast Flow Overlay using Multicast Listener Discovery Protocols", draft-pfister-bier-mld-00 (work in progress), July 2015.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC2291] Slein, J., Vitali, F., Whitehead, E., and D. Durand, "Requirements for a Distributed Authoring and Versioning Protocol for the World Wide Web", RFC 2291, DOI 10.17487/RFC2291, February 1998, <http://www.rfc-editor.org/info/rfc2291>.

[RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <http://www.rfc-editor.org/info/rfc3810>.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <http://www.rfc-editor.org/info/rfc4601>.

[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, DOI 10.17487/RFC4605, August 2006, <http://www.rfc-editor.org/info/rfc4605>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <http://www.rfc-editor.org/info/rfc6514>.

[RFC6826] Wijnands, IJ., Ed., Eckert, T., Leymann, N., and M. Napierala, "Multipoint LDP In-Band Signaling for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6826, DOI 10.17487/RFC6826, January 2013, <http://www.rfc-editor.org/info/rfc6826>.

[RFC6830] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "The Locator/ID Separation Protocol (LISP)", RFC 6830, DOI 10.17487/RFC6830, January 2013, <http://www.rfc-editor.org/info/rfc6830>.

Authors' Addresses

IJsbrand Wijnands
Cisco Systems
De Kleetlaan 6a
Diegem 1831
Belgium

Email: ice@cisco.com

Pierre Pfister
Cisco Systems
Paris
France

Email: pierre.pfister@darou.fr

Jeffrey Zhang
Juniper Networks
10 Technology Park Dr.
Westford MA 01886
US

Email: zzhang@juniper.net

PIM WG
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2017

Stig. Venaas
Cisco Systems, Inc.
Zheng. Zhang
ZTE Corporation

July 7, 2016

PIM IGP EXT
draft-zhang-pim-igp-ext-01

Abstract

This document introduces a method to advertise multicast source information. The information will be flooded all over the network by OSPF, ISIS and Babel extension. This allows PIM Sparse Mode routers with connected receivers to build a Shortest Path Tree straight away, with no need for a shared a tree.

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 8, 2017.

Copyright Notice

Table of Contents

1. Terminology 2
2. Introduction 2
3. Advertisement mechanism 3
4. IGP extension 3
4.1. OSPF extension 3
4.2. ISIS extension 4
4.3. Babel extension 4
5. Security Consideration 4
6. IANA Considerations 5
7. Normative References 5
Authors' Addresses 6

1. Terminology

RP: Rendezvous Point.

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.

MSGI: Multicast Source and group Information as abbreviation.

2. Introduction

[RFC4601] and [RFC7761] introduces that RP can be used to collect the receiver and source information. Obviously, RP may be bottleneck in some busy network. Though the RP-mapping mechanism [RFC6226] is used to make different RP in charge of different groups, it makes the network management more difficult and complex.

[I-D.ietf-pim-source-discovery-bsr] defines an effective way to deliver multicast information by the way of PIM packet flooding. This function is very useful in network with the routers that are all credible and controllable.

Some routers may be attacked or forged in some networks. In these networks, the source information announcement may be forged. There

is authentication method in IGP advertisement, such as OSPF, ISIS and Babel. Authentication can prevent a router from injecting messages with non-existing multicast sources. So the source information announcement may be carried in OSPF, ISIS and Babel extension.

3. Advertisement mechanism

OSPF and ISIS are deployed widely in internet. And the two protocols are the most popular and important routing protocol. The flooding feature is an effective way to advertise the change of network topology. In order to advertise the MSGI, the IGP flooding feature is beneficial to spread the information to PIM routers that have, or potentially may have, connected receivers.

Babel [RFC6126] is a loop-avoiding distance-vector routing protocol that is robust and efficient both in ordinary wired networks and in wireless mesh networks. And multicast service is useful in wired networks and wireless networks. [RFC7298] defines the authentication method of Babel. Babel extension can be used to delivery MSGI.

When a router starts receiving packets from a directly connected source, it should advertise a MSGI for the source in the IGP, and keep doing so as long as the source is active. Along with the IGP flooding, the MSGI will quickly spread all over the network.

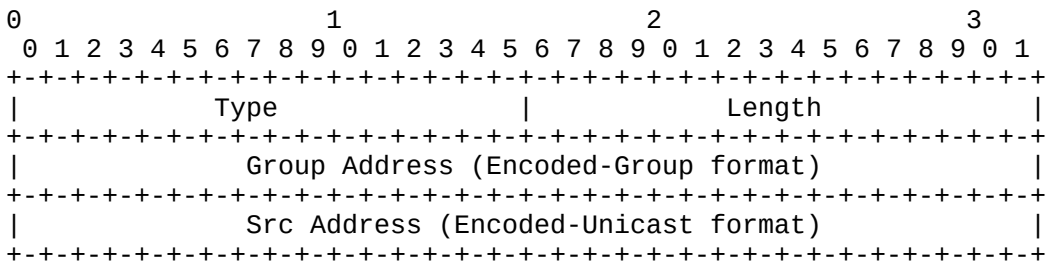
All routers receive the advertisement of the MSGI after flooding. A router that is a LHR, joins the SPT towards the announced source according to standard PIM Sparse Mode procedures, by sending a join to the RPF neighbor towards the source.

Routers that do not have any connected receivers store the MSGI, such that they can immediately join the SPT if they later should become a LHR.

4. IGP extension

4.1. OSPF extension

A new type of the OSPF Opaque LSA is defined for OSPF MSGI capability. And the same for OSPFv2 and OSPFv3. The format is:



- Type : The value is TBD. 12 or later digit can be used.

- 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 [RFC7761].

- Src Address: The source address for the corresponding group. The format for these addresses is given in the Encoded-Unicast address in [RFC7761].

The TLV repeats for many groups and groups. In the case where a source stops sending, the FHR simply stops announcing the TLVs. Then the other routers delete the source information.

4.2. ISIS extension

A new ISIS TLV is defined for the MSGI advertisement. The format of the TLV is same as OSPF.

4.3. Babel extension

A new Babel TLV is defined for MSGI advertisement according to [RFC7557]. The format is same as OSPF.

5. Security Consideration

OSPF and ISIS protocol have the capability of authentication. The security function can be used unchanged for the MSGI advertisement.

The authentication method defined in Babel [RFC7298] can be used unchanged for MSGI advertisement.

6. IANA Considerations

A new OSPF Opaque LSA need to be added for carrying OSPF MSGI TLV.

A new MSGI TLV need to be added for ISIS MSGI advertisement.

A new Babel TLV is defined for MSGI advertisement according to [RFC7557].

7. Normative References

[I-D.ietf-pim-source-discovery-bsr]
Wijnands, I., Venaas, S., Brig, M., and A. Jonasson, "PIM flooding mechanism and source discovery", draft-ietf-pim-source-discovery-bsr-04 (work in progress), March 2016.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <http://www.rfc-editor.org/info/rfc4601>.

[RFC6126] Chroboczek, J., "The Babel Routing Protocol", RFC 6126, DOI 10.17487/RFC6126, April 2011, <http://www.rfc-editor.org/info/rfc6126>.

[RFC6226] Joshi, B., Kessler, A., and D. McWalter, "PIM Group-to-Rendezvous-Point Mapping", RFC 6226, DOI 10.17487/RFC6226, May 2011, <http://www.rfc-editor.org/info/rfc6226>.

[RFC7298] Ovsienko, D., "Babel Hashed Message Authentication Code (HMAC) Cryptographic Authentication", RFC 7298, DOI 10.17487/RFC7298, July 2014, <http://www.rfc-editor.org/info/rfc7298>.

[RFC7557] Chroboczek, J., "Extension Mechanism for the Babel Routing Protocol", RFC 7557, DOI 10.17487/RFC7557, May 2015, <http://www.rfc-editor.org/info/rfc7557>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <http://www.rfc-editor.org/info/rfc7761>.

[RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <http://www.rfc-editor.org/info/rfc7770>.

Authors' Addresses

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

Email: stig@cisco.com

Zheng(Sandy) Zhang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: zhang.zheng@zte.com.cn

PIM WG
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2017

X. Liu
Ericsson
Z. Zhang
ZTE Corporation
A. Peter
Juniper Networks
M. Sivakumar
Cisco Systems
F. Guo
Huawei Technologies
P. McAllister
Metaswitch Networks
July 7, 2016

MSDP YANG module
draft-zhang-pim-msdp-yang-01

Abstract

This document defines a YANG data model for MSDP protocol configuration and operation.

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 8, 2017.

Copyright Notice

Table of Contents

1. Introduction 2
2. Design of the Data Model 2
3. MSDP configuration 5
4. Notifications 5
5. MSDP YANG module 5
6. Contributors 20
7. Normative References 20
Authors' Addresses 21

1. Introduction

[RFC3618] introduces protocol definition of MSDP. This document defines a YANG data model for MSDP. The content is in keeping with [RFC3618].

2. Design of the Data Model

The msdp peer and source-active content is important part of MSDP. The augment should be added below routing-protocol.

```
module: ietf-msdp
augment /rt:routing/rt:control-plane-protocols:
   +--rw msdp!
      +--rw global
      |  +--rw connect-source?   if:interface-ref
      |  +--rw default-peer! {global-default-peer}?
      |  |  +--rw peer-addr        -> ../../../peers/peer/address
      |  |  +--rw prefix-policy?   string {global-default-peer-policy}?
      |  +--rw originating-rp
      |  |  +--rw interface?   if:interface-ref
      |  +--rw sa-filter
      |  |  +--rw in?    string
      |  |  +--rw out?   string
      |  +--rw ttl-threshold?    uint8
      +--rw peers
         +--rw peer* [address]
            +--rw address           inet:ipv4-address
            +--rw authentication
            |  +--rw (authentication-type)?
```

```
            |     +--:(key-chain) {peer-key-chain}?
            |     |  +--rw key-chain?   key-chain:key-chain-ref
            |     +--:(password) {peer-key-chain}?
            |        +--rw key?         string
            +--rw enable?           boolean {peer-admin-enable}?
            +--rw connect-source?   if:interface-ref
            +--rw description?      string {peer-description}?
            +--rw mesh-group?       string
            +--rw peer-as?          string {peer-as}?
            +--rw sa-filter
            |  +--rw in?    string
            |  +--rw out?   string
            +--rw timer
            |  +--rw connect-retry-interval?   uint16 {peer-timer-connect-retry}
?
            |  +--rw holdtime-interval?        uint16 {peer-timer-holdtime}?
            |  +--rw keepalive-interval?       uint16 {peer-timer-keepalive}?
            +--rw ttl-threshold?    uint8
augment /rt:routing-state/rt:control-plane-protocols:
   +--ro msdp!
      +--ro global
      |  +--ro connect-source?   if:interface-ref
      |  +--ro default-peer! {global-default-peer}?
      |  |  +--ro peer-addr        -> ../../../peers/peer/address
      |  |  +--ro prefix-policy?   string {global-default-peer-policy}?
      |  +--ro originating-rp
      |  |  +--ro interface?   if:interface-ref
      |  +--ro sa-filter
      |  |  +--ro in?    string
      |  |  +--ro out?   string
      |  +--ro ttl-threshold?    uint8
      +--ro peers
      |  +--ro peer* [address]
      |     +--ro address                inet:ipv4-address
      |     +--ro authentication
      |     |  +--ro (authentication-type)?
      |     |     +--:(key-chain) {peer-key-chain}?
      |     |     |  +--ro key-chain?   key-chain:key-chain-ref
      |     |     +--:(password) {peer-key-chain}?
      |     |        +--ro key?         string
      |     +--ro enable?                boolean {peer-admin-enable}?
      |     +--ro connect-source?        if:interface-ref
      |     +--ro description?           string {peer-description}?
      |     +--ro mesh-group?            string
      |     +--ro peer-as?               string {peer-as}?
      |     +--ro sa-filter
      |     |  +--ro in?    string
      |     |  +--ro out?   string
      |     +--ro timer
```

```
      |     |  +--ro connect-retry-interval?   uint16 {peer-timer-connect-retry}
?
      |     |  +--ro holdtime-interval?        uint16 {peer-timer-holdtime}?
      |     |  +--ro keepalive-interval?       uint16 {peer-timer-keepalive}?
      |     +--ro ttl-threshold?          uint8
      |     +--ro session-state?          enumeration
      |     +--ro elapsed-time?           uint32
      |     +--ro connect-retry-expire?   uint32
      |     +--ro hold-expire?            uint32
      |     +--ro is-default-peer?        boolean
      |     +--ro keepalive-expire?       uint32
      |     +--ro reset-count?            uint32
      |     +--ro statistics
      |        +--ro discontinuity-time?   yang:date-and-time
      |        +--ro error
      |        |  +--ro rpf-failure?   uint32
      |        +--ro queue
      |        |  +--ro size-in?    uint32
      |        |  +--ro size-out?   uint32
      |        +--ro received
      |        |  +--ro keepalive?      yang:counter64
      |        |  +--ro notification?   yang:counter64
      |        |  +--ro sa-message?     yang:counter64
      |        |  +--ro sa-response?    yang:counter64
      |        |  +--ro sa-request?     yang:counter64
      |        |  +--ro total?          yang:counter64
      |        +--ro sent
      |           +--ro keepalive?      yang:counter64
      |           +--ro notification?   yang:counter64
      |           +--ro sa-message?     yang:counter64
      |           +--ro sa-response?    yang:counter64
      |           +--ro sa-request?     yang:counter64
      |           +--ro total?          yang:counter64
      +--ro sa-cache
         +--ro entry* [group source-addr]
            +--ro group                inet:ipv4-address
            +--ro source-addr          union
            +--ro origin-rp* [rp-address]
            |  +--ro rp-address       inet:ip-address
            |  +--ro is-local-rp?     boolean
            |  +--ro sa-adv-expire?   uint32
            +--ro up-time?             uint32
            +--ro expire?              uint32
            +--ro holddown-interval?   uint32
            +--ro peer-learned-from?   inet:ipv4-address
            +--ro rpf-peer?            inet:ipv4-address
rpcs:
   +---x msdp-clear-peer
   |  +---w input
```

```
|     +---w peer-address?   inet:ipv4-address
+---x msdp-clear-sa-cache {rpc-clear-sa-cache}?
   +---w input
      +---w entry!
      |  +---w group          inet:ipv4-address
      |  +---w source-addr?   union
      +---w peer-address?   inet:ipv4-address
      +---w peer-as?        string
```

3. MSDP configuration

The msdp peers should be configured. And several peers may be in a mesh-group. The Source-Active information may be filtered for peers.

4. Notifications

This part will be updated in later version.

5. MSDP YANG module

```
<CODE BEGINS> file "ietf-msdp@2016-07-06.yang"
module ietf-msdp {
  namespace "urn:ietf:params:xml:ns:yang:ietf-msdp";
  // replace with IANA namespace when assigned
  prefix msdp;

  import ietf-yang-types {
    prefix "yang";
  }

  import ietf-inet-types {
    prefix "inet";
  }

  import ietf-routing {
    prefix "rt";
  }

  import ietf-interfaces {
    prefix "if";
  }

  import ietf-ip {
    prefix "ip";
  }

  import ietf-key-chain {
    prefix "key-chain";
```

```
     }

     organization
       "IETF PIM( Protocols for IP Multicast ) Working Group";

     contact
       "WG Web:   <http://tools.ietf.org/wg/pim/>
        WG List:  <mailto:pim@ietf.org>
        WG Chair: Stig Venaas
                  <mailto:stig@venaas.com>
        WG Chair: Mike McBride
                  <mailto:mmcbride7@gmail.com>

        Editors:   ";

     description
       "The module defines the YANG definitions for MSDP.";

     revision 2016-07-06 {
       description
         "Initial revision.";
       reference
         "RFC XXXX: A YANG Data Model for MSDP.
          RFC 3618: Multicast Source Discovery Protocol (MSDP).
          RFC 4624: Multicast Source Discovery Protocol (MSDP) MIB";
     }

     /*
      * Features
      */
     feature global-connect-source {
       description
         "Support configuration of global connect-source.";
     }

     feature global-default-peer {
       description
         "Support configuration of global default peer.";
     }

     feature global-default-peer-policy {
       description
         "Support configuration of global default peer.";
     }

     feature global-sa-filter {
       description
         "Support configuration of global SA filter.";
```

```
    }

    feature global-ttl-threshold {
      description
        "Support configuration of global ttl-threshold.";
    }

    feature rpc-clear-sa-cache {
      description
        "Support the rpc to clear SA cache.";
    }

    feature peer-admin-enable {
      description
        "Support configuration of peer administrative enabling.";
    }

    feature peer-as {
      description
        "Support configuration of peer AS number.";
    }

    feature peer-connect-source {
      description
        "Support configuration of global connect-source.";
    }

    feature peer-description {
      description
        "Support configuration of peer description.";
    }

    feature peer-key-chain {
      description
        "Support configuration of peer key-chain.";
    }

    feature peer-password {
      description
        "Support configuration of peer key-chain.";
    }

    feature peer-timer-connect-retry {
      description
        "Support configuration of peer timer for connect-retry.";
    }

    feature peer-timer-keepalive {
```

```
      description
        "Support configuration of peer timer for keepalive.";
    }

    feature peer-timer-holdtime {
      description
        "Support configuration of peer timer for holdtime.";
    }

    /*
     * Typedefs
     */

    /*
     * Identities
     */

    /*
     * Groupings
     */
    grouping authentication-container {
      description
        "A container defining authentication attributes.";
      container authentication {
        description
          "A container defining authentication attributes.";
        choice authentication-type {
          case key-chain {
            if-feature peer-key-chain;
            leaf key-chain {
              type key-chain:key-chain-ref;
              description
                "Reference to a key-chain.";
            }
          }
          case password {
            if-feature peer-key-chain;
            leaf key {
              type string;
              description
                "This leaf describes the authentication key.";
            }
            uses key-chain:crypto-algorithm-types;
          }
          description
            "Choice of authentication.";
        }
      }
```

```
    } // authentication-container

    grouping connect-source {
      description "Attribute to configure connect-source.";
      leaf connect-source {
        type if:interface-ref;
        must "/if:interfaces/if:interface[if:name = current()]/"
          + "ip:ipv4" {
          description
            "The interface must have IPv4 enabled.";
        }
        description
          "The interface is to be the source for the TCP connection.
           It is a reference to an entry in the global interface
           list.";
      }
    } // connect-source

    grouping global-config-attributes {
      description "Global MSDP configuration.";

      uses connect-source {
        if-feature global-connect-source;
      }
      container default-peer {
        if-feature global-default-peer;
        presence "";
        description
          "The default peer accepts all MSDP SA messages.
           A default peer is needed in topologies where MSDP peers do
           not coexist with BGP peers. The reverse path forwarding
           (RPF) check on SA messages can fail, and no SA messages are
           accepted. In these cases, you can configure the peer as a
           default peer and bypass RPF checks.";
        leaf peer-addr {
          type leafref {
            path "../../../peers/peer/address";
          }
          mandatory true;
          description
            "Reference to a peer that is in the peer list.";
        }
        leaf prefix-policy {
          if-feature global-default-peer-policy;
          type string;
          description
            "If specified, only those SA entries whose RP is permitted
             in the prefix list are allowed;
```

```
            if not specified, all SA messages from the default peer
            are accepted.";
        }
      } // default-peer

      container originating-rp {
        description
          "The container of originating-rp.";
        leaf interface {
          type if:interface-ref;
          must "/if:interfaces/if:interface[if:name = current()]/"
            + "ip:ipv4" {
            description
              "The interface must have IPv4 enabled.";
          }
          description
            "Reference to an entry in the global interface
             list.
             IP address of the interface is used in the RP field of an
             SA message entry. When Anycast RPs are used, all RPs use
             the same IP address. This parameter can be used to define
             a unique IP address for the RP of each MSDP peer.
             By default, the software uses the RP address of the
             local system.";
        }
      } // originating-rp

      uses sa-filter-container {
        if-feature global-sa-filter;
      }
      uses ttl-threshold {
        if-feature global-ttl-threshold;
      }
    } // global-config-attributes

    grouping global-state-attributes {
      description "Global MSDP state attributes.";
    } // global-state-attributes

    grouping peer-config-attributes {
      description "Per peer configuration for MSDP.";

      uses authentication-container;
      leaf enable {
        if-feature peer-admin-enable;
        type boolean;
        description
          "true to enable peer;
```

```
           false to disable peer.";
       }
       uses connect-source {
         if-feature peer-connect-source;
       }
       leaf description {
         if-feature peer-description;
         type string;
         description
           "The peer description.";
       }
       leaf mesh-group {
         type string;
         description
           "Configure this peer to be a member of a mesh group";
       }
       leaf peer-as {
         if-feature peer-as;
         type string;
         description
           "Peer's autonomous system number (ASN).";
       }
       uses sa-filter-container;
       container timer {
         description "Timer attributes.";
         leaf connect-retry-interval {
           if-feature peer-timer-connect-retry;
           type uint16;
           units seconds;
           default 30;
           description "SHOULD be set to 30 seconds. ";
         }
         leaf holdtime-interval {
           if-feature peer-timer-holdtime;
           type uint16;
           units seconds;
           must ". > 3";
           default 75;
           description "The SA-Hold-Down-Period of this msdp peer.";
         }
         leaf keepalive-interval {
           if-feature peer-timer-keepalive;
           type uint16;
           units seconds;
           must ". > 1 and . < ../holdtime-interval";
           default 60;
           description "The keepalive timer of this msdp peer.";
         }
```

```
      } // timer
      uses ttl-threshold;
    } // peer-config-attributes

    grouping peer-state-attributes {
      description "Per peer state attributes for MSDP.";

      leaf session-state {
        type enumeration {
          enum disabled  {
            description "Disabled.";
          }
          enum inactive {
            description "Inactive.";
          }
          enum listen {
            description "Listen.";
          }
          enum connecting {
            description "Connecting.";
          }
          enum established {
            description "Established.";
          }
        }
        description
          "Peer session state.";
        reference
          "RFC3618: Multicast Source Discovery Protocol (MSDP).";
      }
      leaf elapsed-time {
        type uint32;
        units seconds;
        description "Elapsed time for being in a state.";
      }
      leaf connect-retry-expire {
        type uint32;
        units seconds;
        description "Connect retry expire time of peer connection.";
      }
      leaf hold-expire {
        type uint32;
        units seconds;
        description "Hold expire time of peer connection.";
      }
      leaf is-default-peer {
        type boolean;
        description "If this peer is default peer.";
```

```
      }
      leaf keepalive-expire {
        type uint32;
        units seconds;
        description "Keepalive expire time of this peer.";
      }
      leaf reset-count {
        type uint32;
        description "The reset count of this peer.";
      }
      uses statistics-container;
    } // peer-config-attributes

    grouping sa-cache-state-attributes {
      description "SA cache state attributes for MSDP.";

      leaf up-time {
        type uint32;
        units seconds;
        description "The up time of this sa cache.";
      }
      leaf expire {
        type uint32;
        units seconds;
        description "If this cache has expired.";
      }
      leaf holddown-interval {
        type uint32;
        units seconds;
        description "Holddown timer value for SA forwarding.";
      }
      leaf peer-learned-from {
        type inet:ipv4-address;
        description
          "The address of peer that we learned this SA from .";
      }
      leaf rpf-peer {
        type inet:ipv4-address;
        description "RPF peer.";
      }
    } // sa-cache-state-attributes

    grouping sa-filter-container {
      description "A container defining SA filters.";
      container sa-filter {
        description
          "Specifies an access control list (ACL) to filter source
           active (SA) messages coming in to or going out of the
```

```
           peer.";
        leaf in {
          type string;
          description
            "Filters incoming SA messages only.";
        }
        leaf out {
          type string;
          description
            "Filters outgoing SA messages only.";
        }
      } // sa-filter
    } // sa-filter-container

    grouping ttl-threshold {
      description "Attribute to configure TTL threshold.";
      leaf ttl-threshold {
        type uint8 {
          range 1..255;
        }
        description
          "Maximum number of hops data packets can traverse before
           being dropped.";
      }
    } // sa-ttl-threshold

    grouping statistics-container {
      description
        "A container defining statistics attributes.";
      container statistics {
        description "";
        leaf discontinuity-time {
          type yang:date-and-time;
          description
            "The time on the most recent occasion at which any one
            or more of the statistic counters suffered a
            discontinuity. If no such discontinuities have occurred
            since the last re-initialization of the local
            management subsystem, then this node contains the time
            the local management subsystem re-initialized itself.";
        }
        container error {
          description "";
          uses statistics-error;
        }
        container queue {
          description "";
          uses statistics-queue;
```

```
        }
        container received {
          description "";
          uses statistics-sent-received;
        }
        container sent {
          description "";
          uses statistics-sent-received;
        }
      }
    } // statistics-container

    grouping statistics-error {
      description
        "A grouping defining error statistics
        attributes.";
      leaf rpf-failure {
        type uint32;
        description "";
      }
    } // statistics-error

    grouping statistics-queue {
      description
        "A grouping defining queue statistics
        attributes.";
      leaf size-in {
        type uint32;
        description
          "The size of the input queue.";
      }
      leaf size-out {
        type uint32;
        description
          "The size of the output queue.";
      }
    } // statistics-queue

    grouping statistics-sent-received {
      description
        "A grouping defining sent and received statistics
        attributes.";
      leaf keepalive {
        type yang:counter64;
        description
          "The number of keepalive messages.";
      }
      leaf notification {
```

```
        type yang:counter64;
        description
          "The number of notification messages.";
      }
      leaf sa-message {
        type yang:counter64;
        description
          "The number of SA messages.";
      }
      leaf sa-response {
        type yang:counter64;
        description
          "The number of SA response messages.";
      }
      leaf sa-request {
        type yang:counter64;
        description
          "The number of SA request messages.";
      }
      leaf total {
        type yang:counter64;
        description
          "The number of total messages.";
      }
    } // statistics-sent-received

    /*
     * Configuration data nodes
     */
    augment "/rt:routing/rt:control-plane-protocols" {
      description
        "MSDP augmentation to routing instance configuration.";

      container msdp {
        presence "Container for MSDP protocol.";
        description
          "MSDP configuration data.";

        container global {
          description
            "Global attributes.";
          uses global-config-attributes;
        }

        container peers {
          description
            "Containing a list of peers.";
```

```
          list peer {
            key "address";
            description
              "List of MSDP peers.";
            leaf address {
              type inet:ipv4-address;
              description
                "";
            }
            uses peer-config-attributes;
          } // peer
        } // peers
      } // msdp
    } // augment

    /*
     * Operational state data nodes
     */
    augment "/rt:routing-state/rt:control-plane-protocols" {
      description
        "MSDP augmentation to routing instance state.";

      container msdp {
        presence "Container for MSDP protocol.";
        description
          "MSDP state data.";

        container global {
          description
            "Global attributes.";
          uses global-config-attributes;
          uses global-state-attributes;
        }

        container peers {
          description
            "Containing a list of peers.";

          list peer {
            key "address";
            description
              "List of MSDP peers.";
            leaf address {
              type inet:ipv4-address;
              description
                "The address of peer";
            }
            uses peer-config-attributes;
```

```
            uses peer-state-attributes;
          } // peer
        } // peers

        container sa-cache {
          description
            "The sa cache information.";
          list entry {
            key "group source-addr";
            description "";
            leaf group {
              type inet:ipv4-address;
              description "The group address of this sa cache.";
            }
            leaf source-addr {
              type union {
                type enumeration {
                  enum '*' {
                    description "The source addr of this sa cache.";
                  }
                }
                type inet:ipv4-address;
              }
              description "";
            }
            list origin-rp {
              key "rp-address";
              description
                "";
              leaf rp-address {
                type inet:ip-address;
                description "The rp address.";
              }
              leaf is-local-rp {
                type boolean;
                description "";
              }
              leaf sa-adv-expire {
                type uint32;
                units seconds;
                description
                  "Periodic SA advertisement timer exipiring time on
                   a local RP.";
              }
            }
            uses sa-cache-state-attributes;
          } // entry
        } // sa-cache
```

```
      } // msdp
    } // augment

    /*
     * RPCs
     */
    rpc msdp-clear-peer {
      description
        "Clears the session to the peer.";
      input {
        leaf peer-address {
          type inet:ipv4-address;
          description
            "Address of peer to be cleared. If this is not provided
             then all peers are cleared.";
        }
      }
    }

    rpc msdp-clear-sa-cache {
      if-feature rpc-clear-sa-cache;
      description
        "Clears MSDP source active (SA) cache entries.";
      input {
        container entry {
          presence "";
          description
            "The SA cache (S,G) or (*,G) entry to be cleared. If this
             is not provided, all entries are cleared.";
          leaf group {
            type inet:ipv4-address;
            mandatory true;
            description "";
          }
          leaf source-addr {
            type union {
              type enumeration {
                enum '*' {
                  description "";
                }
              }
              type inet:ipv4-address;
            }
            description "";
          }
        } // s-g
        leaf peer-address {
          type inet:ipv4-address;
```

```
            description
              "Peer IP address from which MSDP SA cache entries have been
               learned. If this is not provided, entries learned from all
               peers are cleared.";
          }
          leaf peer-as {
            type string;
            description
              "ASN from which MSDP SA cache entries have been learned.
               If this is not provided, entries learned from all AS's
               are cleared.";
          }
        }
      }

      /*
       * Notifications
       */
    }
    <CODE ENDS>
```

6. Contributors

The authors would like to thank Yisong Liu (liuyisong@huawei.com), Benchong Xu (xu.benchong@zte.com.cn), Tanmoy Kundu (tanmoy.kundu@alcatel-lucent.com) for their valuable contributions.

7. Normative References

[I-D.ietf-netmod-routing-cfg] Lhotka, L. and A. Lindem, "A YANG Data Model for Routing Management", draft-ietf-netmod-routing-cfg-22 (work in progress), July 2016.

[RFC3618] Fenner, B., Ed. and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", RFC 3618, DOI 10.17487/RFC3618, October 2003, <http://www.rfc-editor.org/info/rfc3618>.

[RFC4624] Fenner, B. and D. Thaler, "Multicast Source Discovery Protocol (MSDP) MIB", RFC 4624, DOI 10.17487/RFC4624, October 2006, <http://www.rfc-editor.org/info/rfc4624>.

[RFC6087] Bierman, A., "Guidelines for Authors and Reviewers of YANG Data Model Documents", RFC 6087, DOI 10.17487/RFC6087, January 2011, <http://www.rfc-editor.org/info/rfc6087>.

Authors' Addresses

Xufeng Liu
Ericsson
1595 Spring Hill Road, Suite 500
Vienna VA 22182
USA

Email: xliu@kuatrotech.com

Zheng(Sandy) Zhang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: zhang.zheng@zte.com.cn

Anish Peter
Juniper Networks
Electra, Exora Business Park
Bangalore, KA 560103
India

Email: anishp@juniper.net

Mahesh Sivakumar
Cisco Systems
510 McCarthy Boulevard
Milpitas, California
USA

Email: masivaku@cisco.com

Feng Guo
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: guofeng@huawei.com

Pete McAllister
Metaswitch Networks
100 Church Street
Enfield EN2 6BQ
UK

Email: pete.mcallister@metaswitch.com