BIER Working Group   IJ. Wijnands
Internet-Draft   P. Pfister
Intended status: Standards Track   Cisco Systems
Expires: January 9, 2017   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

Copyright Notice

Table of Contents

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.

```
                       +---- MOSP ----+
                                           LAN2
                                    ( R3 ) -|
                  LAN1             /        |
              S H1-|-( R1 )--( R2 )         |- H2  (joined G)
                                   \        |
                                    ( R4 ) -|
```

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:

```
                       +---- MOSP ----+
                                           LAN2
                                    ( R3 ) -|
                  LAN1             /        |
              S H1-|-( R1 )--( R2 )         |- H2  (joined G)
                                   \        |
                                    ( R4 ) -|
                                       |
                                       - LAN3
                                       |
                                       H3 (joined G)
```

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.

```
                        +---- MOSP ----+
                 LAN1
                  |- ( R1 )
                  |        \               LAN2
            S H1 -|        ( R3 ) -- ( R4 ) -|- H2  (joined G)
                  |        /
                  |- ( R2 )
```

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.

o The active DA expires.

o 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