PIM WG   Zheng. Zhang
Internet-Draft   Fangwei. Hu
Intended status: Standards Track   Benchong. Xu
Expires: July 6, 2019   ZTE Corporation
                                             Mankamana. Mishra
                                                 Cisco Systems
                                               January 2, 2019

PIM DR Improvement
draft-ietf-pim-dr-improvement-06.txt

Abstract

   PIM is widely deployed multicast protocol. PIM protocol is defined
   in [RFC7761]. As deployment for PIM protocol is growing day by day,
   user expects lower traffic loss and faster convergence in case of any
   network failure. This document provides extension to the existing
   protocol which would improve stability of PIM protocol with respect
   to traffic loss and convergence time when the PIM DR is down.

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 https://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 July 6, 2019.

Table of Contents

1. Introduction

   Multicast technology is used widely. Many modern technologies, such
   as IPTV, Net-Meeting, use PIM-SM to facilitate multicast service.
   There are many events that will influence the quality of multicast
   services. Like the change of unicast routes, the change of the PIM-
   SM DR may cause the loss of multicast packets too.

   After a DR on a shared-media LAN went down, other routers will elect
   a new DR after the expiration of Hello-Holdtime. The default value
   of Hello-Holdtime is 105 seconds. Although the minimum Hello
   interval can be adjust to 1 second and the Hello-Holdtime is 3.5
   times Hello interval. Thus, the detection of DR Down event cannot be
   guaranteed in less than 3.5 seconds. And it is still too long for
   modern multicast services. Still, may multicast packets will be
   lost. The quality of IPTV and Net-Meeting will be influenced.

```
              \                                    /
               \                                  /
           -------                            -------
           |  A  |                            |  B  |
           -------                            -------
              | DR                               |
              |                                  |
           -------                            -------
           | SW  |----------------------------| SW  |
           -------                            -------
              |                                  |
```
                    Figure 1: An example of multicast network

   For example, there are two routers on one Ethernet. RouterA is
   elected to DR. When RouterA is down, the multicast packets are
   discarded until the RouterB is elected to DR and RouterB imports the
   multicast flows successfully.

   We suppose that there is only a RouterA in the Ethernet at first in
   Figure 1. RouterA is the DR which is responsible for forwarding
   multicast flows. When RouterB connects to the Ethernet segment,
   RouterB will be elected as DR because of its higher priority. So
   RouterA will stop forwarding multicast packets. The multicast flows
   will not recover until RouterB pulls the multicast flows after it is
   elected to DR.

   So if we want to increase the stability of DR, carrying DR/ BDR role
   information in PIM hello packet is a feasible way to show the DR/ BDR
   roles explicitly. It avoids the confusion caused by new comers which
   has a higher priority.

2. Terminology

   Backup Designated Router (BDR): Like DR, A BDR which acts on behalf
   of directly connected hosts in a shared-media LAN. But BDR MUST not
   forward the flows when DR works normally. When DR is down, the BDR
   will forward multicast flows immediately. A single BDR MUST be
   elected per interface like the DR.

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

3. PIM hello message format

   In [RFC7761], the PIM hello message format is defined. In this
   document, we define two new option values which are including Type,
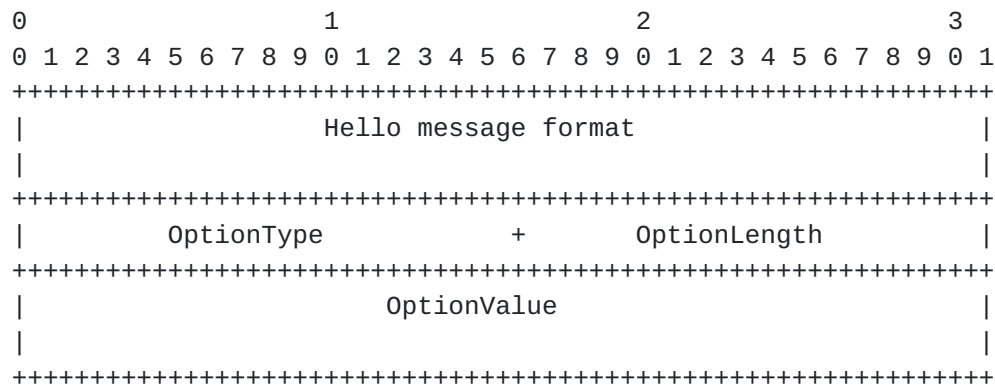   Length, and Value.

```
           0                   1                   2                   3
           0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
          +++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
          |                   Hello message format                       |
          |                                                              |
          +++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
          |         OptionType            +       OptionLength           |
          +++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
          |                       OptionValue                            |
          |                                                              |
          +++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
                       Figure 2: Hello message format
```

3.1. DR Address Option format

o OptionType : The value is TBD.

o OptionLength: If the network is support IPv4, the OptionLength is
   4 octets. If the network is support IPv6, the OptionLength is 16
   octets.

o OptionValue: The OptionValue is IP address of DR. If the network
   is support IPv4, the value is IPv4 address of DR. If the network
   is support IPv6, the value is IPv6 address of DR.

3.2. BDR Address Option format

o OptionType : The value is TBD.

o OptionLength: If the network is support IPv4, the OptionLength is
   4 octets. If the network is support IPv6, the OptionLength is 16
   octets.

o OptionValue: The OptionValue is IP address of BDR. If the network
   is support IPv4, the value is IPv4 address of BDR. If the network
   is support IPv6, the value is IPv6 address of BDR.

4. The Protocol Treatment

Carrying DR/ BDR role information in PIM hello packet is a feasible
way to keep the stability of DR. It avoid the confusion caused by
new comers which has a higher priority. So there are some changes in
PIM hello procedure and interface state machine.

A new router starts to send hello messages with the values of DR and
BDR are all set to 0 after its interface is enabled in PIM on a
shared-media LAN. When the router receives hello messages from other
routers on the same shared-media LAN, the router will check if the

value of DR is filled. If the value of DR is filled with IP address
of router which is sending hello messages, the router will store the
IP address as the DR address of this interface.

Then the new router compares the priority and IP address itself to
the stored information of DR and BDR according to the algorithm of
[RFC7761]. If the new router notices that it is better to be DR than
the current DR or BDR. The router will make itself the BDR, and send
new hello messages with its IP address as BDR and current DR. If the
router notices that the current DR has the highest priority in the
shared-media LAN, but the current BDR is set to 0x00000000 if IPv4
addresses are in use or 0:0:0:0:0:0:0:0/128 if IPv6 addresses are in
use in the received hello messages (To be simplify, we use 0x0 in
abbreviation in following parts of the draft), or the current BDR is
not better than the new router, the new router will elect itself to
BDR. If the router notices that it is not better to be DR than
current DR and BDR, the router will follow the current DR and BDR.

When the new router becomes the new BDR, the router will join the
current multicast groups, import multicast flows from upstream
routers. But the BDR MUST not forward the multicast flows to avoid
the duplicate multicast packets in the shared-media LAN. The new
router will monitor the DR. The method that BDR monitors the DR may
be BFD technology or other ways that can be used to detect link/node
failure quickly. When the DR becomes unavailable because of the down
or other reasons, the BDR will forward multicast flows immediately.

4.1. Deployment Choice

DR / BDR election SHOULD be handled in two ways. Selection of which
procedure to use would be totally dependent on deployment scenario.

1. The algorithm defined in [RFC7761] should be used if it is ok to
adopt with new DR as and when they are available, and the loss caused
by DR changing is acceptable.

2. If the deployment requirement is to have minium packets loss when
DR changing the mechanism defined in this draft should be used. That
is, if the new router notices that it is better to be DR than the
current DR or BDR, the router will make itself the BDR, and send new
hello message with its IP address as BDR and current DR.

According to section 4.9.2 defined in [RFC7761], the device receives
unknown options Hello packet will ignore it. So the new extension
defined in this draft will not influence the stability of neighbor.
But if the router which has the ability defined in this draft
receives non-DR/BDR capable Hello messages defined in [RFC7761], the
router MAY stop sending DR/BDR capable Hello messages in the LAN and

go back to use the advertisement and election algorithm defined in
[RFC7761].

4.2. Election Algorithm

The DR and BDR election is according the rules defined below, the
algorithm is similar to the DR election definition in [RFC2328].

(1) Note the current values for the network's Designated Router and
Backup Designated Router. This is used later for comparison
purposes.

(2) Calculate the new Backup Designated Router for the network as
follows. The router that has not declared itself to be Designated
Router is eligible to become Backup Designated Router. The one which
have the highest priority will be chosen to be Backup Designed
Router. In case of a tie, the one having the highest Router ID is
chosen.

(3) Calculate the new Designated Router for the network as follows.
If one or more of the routers have declared themselves Designated
Router (i.e., they are currently listing themselves as Designated
Router in their Hello Packets) the one having highest Router Priority
is declared to be Designated Router. In case of a tie, the one
having the highest Router ID is chosen. If no routers have declared
themselves Designated Router, assign the Designated Router to be the
same as the newly elected Backup Designated Router.

(4) If Router X is now newly the Designated Router or newly the
Backup Designated Router, or is now no longer the Designated Router
or no longer the Backup Designated Router, repeat steps 2 and 3, and
then proceed to step 5. For example, if Router X is now the
Designated Router, when step 2 is repeated X will no longer be
eligible for Backup Designated Router election. Among other things,
this will ensure that no router will declare itself both Backup
Designated Router and Designated Router.

(5) As a result of these calculations, the router itself may now be
Designated Router or Backup Designated Router.

The reason behind the election algorithm's complexity is the desire
for the DR stability.

The above procedure may elect the same router to be both Designated
Router and Backup Designated Router, although that router will never
be the calculating router (Router X) itself. The elected Designated
Router may not be the router having the highest Router Priority. If
Router X is not itself eligible to become Designated Router, it is

possible that neither a Backup Designated Router nor a Designated
Router will be selected in the above procedure. Note also that if
Router X is the only attached router that is eligible to become
Designated Router, it will select itself as Designated Router and
there will be no Backup Designated Router for the network.

4.3. Sending Hello Messages

According to Section 4.3.1 in [RFC7761], when a new router's
interface is enabled in PIM protocol, the router sends Hello messages
with the values of DR and BDR are filled with 0x0. Then the
interface is in Waiting state and start the hold-timer which is equal
to the Neighbor Liveness Timer. When the timer is expired, the
interface will elect the DR and BDR according to the DR election
rules.

When a new router sets itself BDR after receive hello messages from
other routers, the router send hello messages with the value of DR is
set to the IP address of current DR and the value of BDR is set to
the IP address of the router itself.

A current BDR MUST set itself DROther after it receives Hello
messages from other routers, the router will send hello messages with
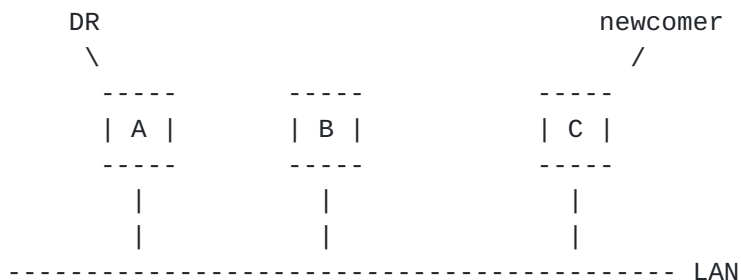the value of DR is set to current DR and the value of BDR is set to
new BDR.

```
            DR                          newcomer
             \                             /
          -----        -----        -----
          | A |        | B |        | C |
          -----        -----        -----
            |            |            |
            |            |            |
        ------------------------------------------- LAN
                       Figure 3
```

For example, there is a stable LAN that includes RouterA and RouterB.
RouterA is the DR which has the best priority. RouterC is a
newcomer. RouterC sends hello packet with the DR and BDR is all set
to zero.

If RouterC cannot send hello packet with the DR/BDR capability,
Router C MAY send the hello packet according the rule defined in
[RFC7761].

If deployment requirement is to adopt with new DR as and when they
are available, a new router with highest priority or best IP address
sends hello packet with DR and BDR all set to zero at first. It

   sends hello packet with itself set to DR after it finish join all the
   existing multicast groups. Then current DR compares with the new
   router, the new router will be final DR.

4.4. Receiving Hello Messages

   When the values of DR and BDR which are carried by hello messages are
   received is all set to 0x0, the router MUST elect the DR using
   procedure defined in DR election algorithm after the hold-timer
   expires. And elect a new BDR which is the best choice except DR.
   The election cases can be executed as following:

   In case the value of DR which is carried by received hello messages
   is not 0x0, and the value of BDR is set to 0x0, when the hold-timer
   expires there is no hello packet from other router is received, the
   router will elect itself to BDR.

   In case either of the values of DR and BDR that are carried by
   received hello messages are larger than 0x0. The router will mark
   the current DR, and compare itself and the BDR in message. When the
   router notice that it is better to be DR than current BDR. The
   router will elect itself to the BDR.

   When a router receives a new hello message with the values of DR and
   BDR are set to 0x0. The router will compare the new router with
   current information. If the router noticed that the new router is
   better to be DR than itself, or the new router is better to be BDR
   than the current BDR, the router will set the BDR to the new router.

   When current DR receives hello packet with DR set larger than zero,
   algorithm defined in section 4.1 can be used to select the final DR.

   As illustrated in Figure 3, after RouterC sends hello packet, RouterC
   will not elect the DR until hold-timer expired. During the period,
   RouterC should receive the hello packets from RouterA and RouterB.
   RouterC accepts the result that RouterA is the DR. In case RouterC
   has the lowest priority than RouterA and RouterB, RouterC will also
   accept that Router B is the BDR. In case RouterC has the
   intermediate priority among the three routers, RouterC will treat
   itself as new BDR after the hold-timer expired. In case RouterC has
   the highest priority among the three routers, RouterC will treat
   RouterA which is the current DR as DR, and RouterC will treat itself
   as new BDR. If the network administrator thinks that RouterC should
   be new DR, the DR changing should be triggered manually.

   Exception: During the hold-timer period, RouterC receives only the
   hello packet from RouterA. When the hold-timer expired, RouterC
   treats RouterA as DR. and RouterC treats itself as BDR. In case

RouterC only receives the hello packet from RouterB during the hold-
timer period, RouterC will compare the priority between RouterB and
itself to elect the new DR. In these situations, some interfaces or
links go wrong in the LAN.

4.5. The treatment

When all the routers on a shared-media LAN are start to work on the
same time, the election result of DR is same as [RFC7761]. And all
the routers will elect a BDR which is next best to DR. The routers
in the network will store the DR and BDR. The hello messages sent by
all the routers are same with the value of DR and BDR are all set.

When a new router start to work on a shared-media LAN and receive
hello messages from other routers that the value of DR is set. The
new router will not change the current DR even if it is superior to
the current DR. If the new router is superior to current BDR, the
new router will replace the current BDR.

When the routers receive hello message from a new router, the routers
will compare the new router and all the other routers on the LAN. If
the new router is superior to current BDR, the new router will be new
BDR. Then the old BDR will send prune message to upstream routers.

As a result, the BDR is the one which has the highest priority except
DR. Once the DR is elected, the DR will not change until it fails or
manually adjustment. After the DR and BDR are elected, the routers
in the network will store the address of DR and BDR.

4.6. Sender side

DR/BDR function also can be used in source side that multiple routers
and source is in same shared-media network. The algorithm is the
same as the receiver side. Only the BDR need not build multicast
tree from downstream router.

5. Compatibility

If the LAN is a hybrid network that there are some routers which have
DR/BDR capability and the other routers which have not DR/BDR
capability. All the routers MAY go backward to use the algorithm
defined in [RFC7761].

6. Deployment suggestion

If there are two and more routers which is responsible for multicast
flow forwarding on a shared-media LAN, and the multicast services is

sensitive to the lost of multicast packets, the function of DR and
BDR defined in this document SHOULD be deployed.

7. Security Considerations

For general PIM Security Considerations.

8. IANA Considerations

IANA is requested to allocate OptionType in TLVs of hello message.
Include DR and BDR.

9. Acknowledgements

The authors would like to thank Greg Mirsky for their valuable
comments and suggestions.

10. Normative References

[HRW]   IEEE, "Using name-based mappings to increase hit rates",
            IEEE HRW, February 1998.

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

[RFC2362] Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering,
            S., Handley, M., Jacobson, V., Liu, C., Sharma, P., and L.
            Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM):
            Protocol Specification", RFC 2362, DOI 10.17487/RFC2362,
            June 1998, <https://www.rfc-editor.org/info/rfc2362>.

[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>.

Authors' Addresses

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

   Email: zhang.zheng@zte.com.cn

Fangwei Hu
ZTE Corporation
No.889 Bibo Rd
Shanghai
China

Email: hu.fangwei@zte.com.cn

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