PIM WG   Z. Zhang
Internet-Draft   ZTE Corporation
Intended status: Standards Track   F. Hu
Expires: August 21, 2021   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.

Status of This Memo

Copyright Notice

Table of Contents

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.
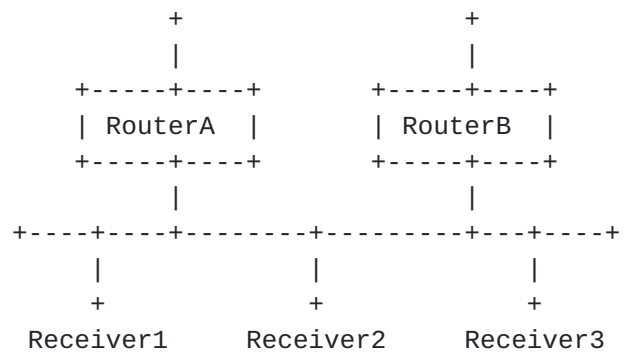
```
                    +                   +
                    |                   |
               +-----+----+        +-----+----+
               | RouterA  |        | RouterB  |
               +-----+----+        +-----+----+
                     |                   |
             +----+----+--------+--------+---+----+
                  |             |            |
                  +             +            +
               Receiver1     Receiver2    Receiver3
             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].

o 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.

o 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.

o 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:

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

o 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:

   o 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.

   o 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.
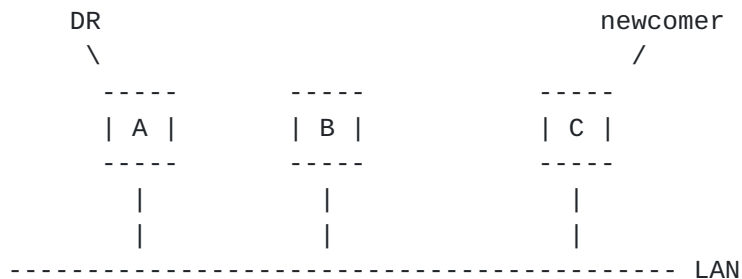
```
          DR                              newcomer
           \                                /
          -----        -----            -----
          | A |        | B |            | C |
          -----        -----            -----
            |            |                |
            |            |                |
        ----------------------------------------- LAN
                      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].

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            OptionType          |           OptionLength        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           OptionValue                         |
|                             ...                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                  Figure 3: Hello Option Format
```

4.1. DR Address Option format

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type = 37          |       Length = <Variable>     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             DR Address  (Encoded-Unicast format)              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                  Figure 4: DR Address Option
```
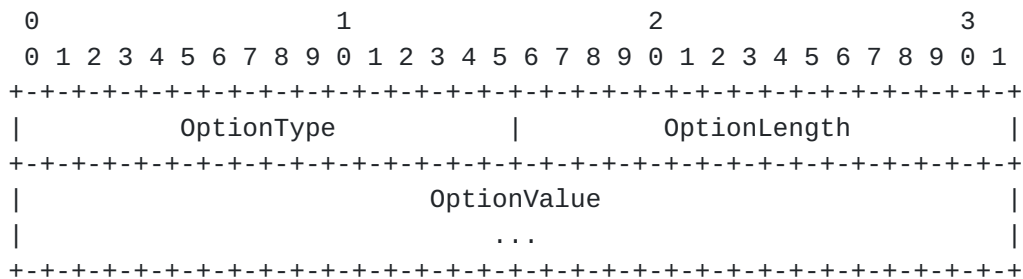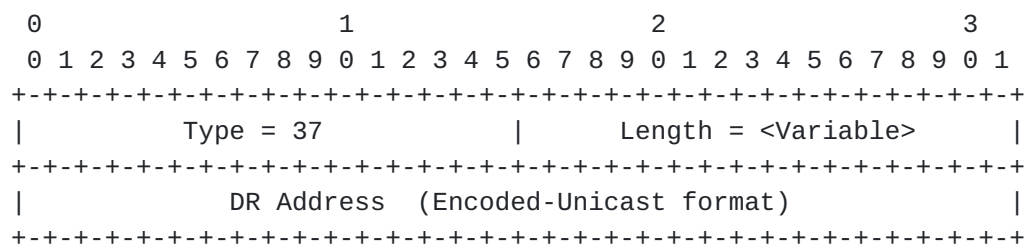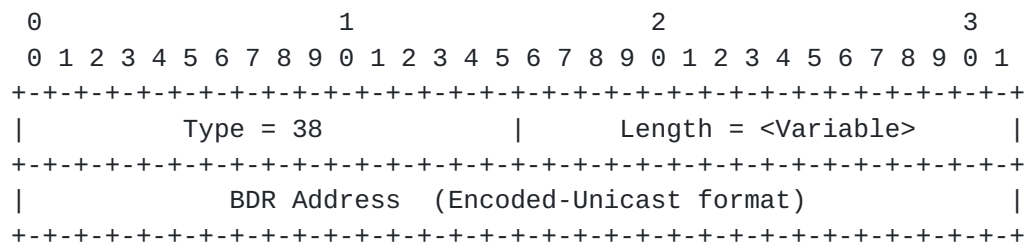
   o OptionType : The value is 37.

   o OptionLength: 4 bytes if using IPv4 and 16 bytes if using IPv6.

   o 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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type = 38          |       Length = <Variable>     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             BDR Address  (Encoded-Unicast format)             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                  Figure 5: BDR Address Option
```

   o OptionType : The value is 38.

   o OptionLength: 4 bytes if using IPv4 and 16 bytes if using IPv6.

   o 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