PIM WG
Internet-Draft
Intended status: Standards Track
Expires: February 16, 2020

Zheng. Zhang
ZTE Corporation
Fangwei. Hu
Individual
Benchong. Xu
ZTE Corporation
Mankamana. Mishra
Cisco Systems
August 15, 2019

PIM DR Improvement
draft-ietf-pim-dr-improvement-08.txt

Abstract

Protocol Independent Multicast - Sparse Mode (PIM-SM) is widely deployed multicast protocol. 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 an extension to the existing protocol which would improve the stability of the PIM protocol with respect to traffic loss and convergence time when the PIM DR role changes.

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 February 16, 2020.

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 goes 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 adjusted to 1 second, 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 too long for modern multicast services. Still, many multicast packets will be lost. The quality of IPTV and Net-Meeting will be influenced.

Figure 1: An example of multicast network

For example, there are two routers on one LAN. Two SWs (Layer 2 switches) provide shared-media LAN connection. RouterA is elected as DR. When RouterA goes 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 LAN at first in Figure 1. RouterA is the DR which is responsible for forwarding multicast flows. When RouterB connects to the LAN, RouterB will be elected as DR because of its higher priority. 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 newcomers which have a higher priority.

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

Backup Designated Router (BDR): Like DR (Designated Router), 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 goes down, the BDR will forward multicast flows immediately. A single BDR MUST be elected per interface like the DR.

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

3. PIM hello message format

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

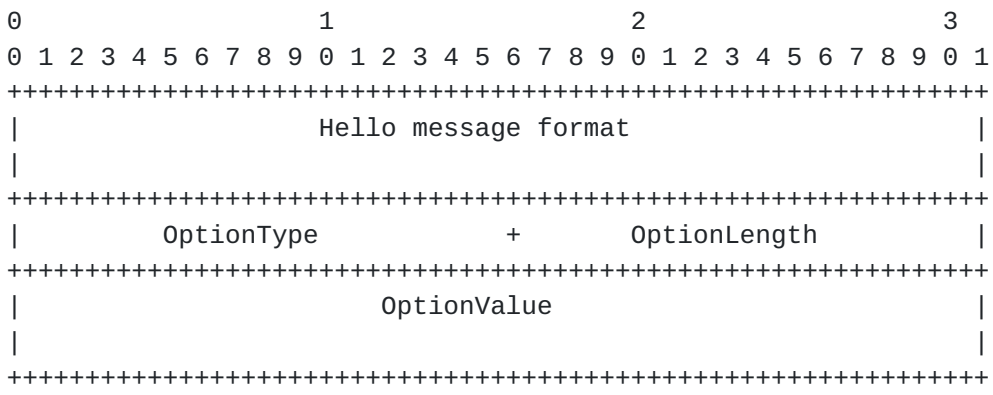

Figure 2: Hello message format

3.1. DR Address Option format

- OptionType : The value is TBD1.
- OptionLength: If the IP version of PIM message is IPv4, the OptionLength is 4 octets. If the IP version of PIM message is IPv6, the OptionLength is 16 octets.
- OptionValue: The OptionValue is IP address of DR. If the IP version of PIM message is IPv4, the value is the IPv4 address of DR. If the IP version of PIM message is IPv6, the value is the IPv6 address of DR.

3.2. BDR Address Option format

- OptionType : The value is TBD2.
- OptionLength: If the IP version of PIM message is IPv4, the OptionLength is 4 octets. If the IP version of PIM message is IPv6, the OptionLength is 16 octets.
- OptionValue: The OptionValue is IP address of BDR. If the IP version of PIM message is IPv4, the value is the IPv4 address of BDR. If the IP version of PIM message is IPv6, the value is the IPv6 address of BDR.

4. Protocol Specification

Carrying DR/ BDR role information in PIM hello packet is a feasible way to keep the stability of DR. It avoids the confusion caused by newcomers which have 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 the 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 new 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 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 in the received hello messages (To 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, and 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 Bidirectional Forwarding Detection (BFD) for Multi-point Networks and Protocol Independent Multicast [I-D.ietf-pim-bfd-p2mp-use-case] technology, BFD (Bidirectional Forwarding Detection) [RFC5880] 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.

BFD for PIM function defined in [I-D.ietf-pim-bfd-p2mp-use-case], or asynchronous mode defined in BFD [RFC5880] are suggested to be used for the DR failure detection. BDR monitors DR after the BFD session between DR and BDR is established. For example, an aggressive BFD session that achieves a detection time of 300 milliseconds, by using a transmit interval of 100 milliseconds and a detect multiplier of 3. So BDR can replace DR to forward flows when DR goes down within sub

second. The other BFD modes can also be used to monitor the failure of DR, the network administrator should choose the most suitable function.

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 minimum 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 new 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 to the rules defined below, the algorithm is similar to the DR election defined 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 has the highest priority will be chosen to be Backup Designated Router. In case of a tie, the one having the highest primary address 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 primary address 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 starts 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 receives hello messages from other routers, the router sends 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 router which is eligible to be BDR/DR, the router

will send hello messages with the value of DR is set to current DR and the value of BDR is set to the new BDR.

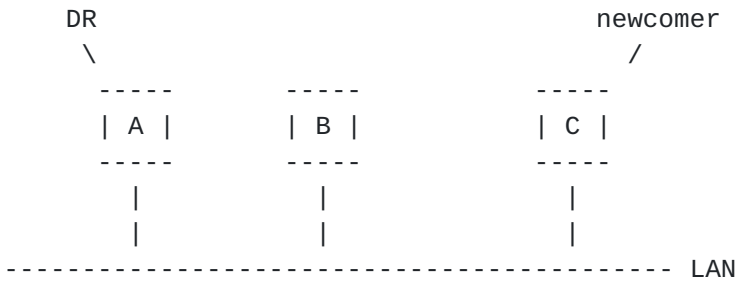

Figure 3

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

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

If deployment requirement is to adopt with a new DR when it is available, a new router with the highest priority or the highest IP address sends hello packet with DR and BDR are 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 the final DR.

4.4. Receiving Hello Messages

When the values of DR and BDR which are carried by hello messages received are 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 follows:

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 is greater than 0x0. The router will mark the current DR, and compare itself with the BDR in the message. When the router notices that it is better to be DR than the 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 the value of DR is set larger than zero, the algorithm defined in section 4.2 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 the new BDR. If the network administrator thinks that RouterC should be the new DR, the DR changing should be triggered manually. That is RouterC will be elected as DR after it sends hello message with DR is set to RouterC itself.

Exception: In case RouterC receives only the hello packet from RouterA during the hold-timer period, 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

If all the routers on a shared-media LAN have started working at the same time, then the election result of DR is same as the definition in [RFC7761]. And all the routers will elect a BDR which is next best to DR. The routers in the network MUST store the DR and BDR. The hello messages sent by all the routers are the same with the value of DR and BDR are all set. When a new router is activated on the shared-media LAN and receives hello messages from other routers with the value of DR is already 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 a hello message from a new router, the routers compare the new router and all the other routers on the LAN. If the new router is superior to the current BDR, the new router will be the new BDR. Then the "old" BDR will send the Prune message to upstream routers.

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

4.6. Sender side

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

5. Compatibility

If the LAN is a hybrid network that there are some routers which support DR/BDR capability and the other routers which do not support DR/BDR capability. All the routers MUST go backward to use the election algorithm defined in [RFC7761]. And the values of DR and BDR carried in hello message MUST be set to zero. That is once a router sends hello messages with no DR/BDR options, the DR election MUST go backward to the definition in [RFC7761].

If the routers find that all the routers in the LAN support DR/BDR capability by the hello messages with DR/BDR options set, they MUST elect DR and BDR according the algorithm defined in this document. And the routers MUST send hello messages with correct DR/BDR options set.

In case there is only one router which does not support DR/BDR capability in a shared-media LAN, the other routers in the LAN send hello messages with the values of DR and BDR are set to zero, the router which does not support DR/BDR capability ignores the options. All the routers elect DR according to the algorithm defined in [RFC7761]. When the router which does not support DR/BDR capability goes away, the routers in the LAN MUST elect DR/BDR according to the algorithm defined in this document, and send hello messages with correct DR/BDR options set.

This draft allows DR election to be sticky by not unnecessarily changing the DR when routers go down or come up. This is done by introducing new PIM Hello options. Both this draft, and the draft [I-D.mankamana-pim-bdr], introduce a backup DR. The latter draft

does this without introducing new options, but does not consider the sticky behavior.

6. Security Considerations

If an attacker which has the highest priority participates in the DR election when a shared-media LAN starts to work, it will be elected as DR, but it may not forward flows to receivers. And the attacker remains DR position even if a legal router which has a higher priority joins the LAN.

If an attacker is a newcomer which has a higher priority than the existed BDR, it will be elected as the new BDR, but it may not monitor DR, import multicast flows and forward flows to receiver when DR is down.

In order to avoid these situations, source authentication should be used to identify the validity of the DR/BDR candidates. Authentication methods mentioned in section 6 RFC7761 can be used.

And the network administrator should consider the potential BFD session attack if BFD is used between BDR and DR for DR failure detection. The security function mentioned in section 9 RFC5880 can be used.

7. IANA Considerations

IANA is requested to allocate two OptionTypes in TLVs of hello message: DR Address Option and BDR Address Option. The strings TBD1 and TBD2 will be replaced by the assigned values.

8. Acknowledgements

The authors would like to thank 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>.

[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>.

[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>.

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-02 (work in progress), July 2019.

[I-D.mankamana-pim-bdr]
mishra, m., "PIM Backup Designated Router Procedure", draft-mankamana-pim-bdr-02 (work in progress), April 2019.

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

[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