Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 17, 2012

Yiqun Cai
Sri Vallepalli
Heidi Ou
Cisco Systems, Inc.
Andy Green
British Telecom

October 15, 2011

Protocol Independent Multicast DR Load Balancing
draft-hou-pim-drlb-00.txt

Abstract

On a multi-access network such as an Ethernet, one of the PIM routers is elected as a Designated Routers (DR). The PIM DR has two roles in the PIM protocol. On the first hop network, the PIM DR is responsible for registering an active source to the RP if the group is operated in PIM SM. On the last hop network, the PIM DR is responsible for tracking local multicast listeners and forwarding traffic to these listeners if the group is operated in PIM SM/SSM/DM. In this document, we propose a modification to the PIM protocol that allows multiple of these last hop routers to be selected so that the forwarding load can be distributed to and handled among these routers. A router responsible for forwarding for a particular group is called a Group Designated Router (GDR).

Status of this Memo

Copyright Notice

Table of Contents

1. Terminologies

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].

With respect to PIM, this document follows the terminology that has been defined in [RFC4601].

This document also introduces the following new acronyms:

- GDR: GDR stands for "Group Designated Router". For each multicast group, a hash algorithm (described below) is used to select one of the routers as GDR. The GDR is responsible for initiating the forwarding tree building for the corresponding group.

- GDR Candidate: a last hop router that has potential to become a GDR. A GDR Candidate must have the same DR priority as the DR router. It must send and process received new PIM Hello Options as defined in this document. There might be more than one GDR candidate on a LAN. But only one can become GDR for a specific multicast group.

2. Introduction

On a multi-access network such as an Ethernet, one of the PIM routers is elected as a Designated Routers (DR). The PIM DR has two roles in the PIM protocol. On the first hop network, the PIM DR is responsible for registering an active source to the RP if the group is operated in PIM SM. On the last hop network, the PIM DR is responsible for tracking local multicast listeners and forwarding to these listeners if the group is operated in PIM SM/SSM/DM.

Considering the following last hop network in Figure 1.

```
                        ( core networks )
                          |     |     |
                          |     |     |
                         R1    R2     R3
                          |     |     |
                       --(last hop LAN)--
                                |
                                |
                        (many receivers)
```

Figure 1: Last Hop Network

Assuming R1 is elected as the Designated Router. According to [RFC4601], R1 will be responsible for forwarding to the last hop LAN. In addition to keeping track of IGMP and MLD membership reports, R1 is also responsible for initiating the creation of source and/or shared trees towards the senders or the RPs.

Forcing sole data plane forwarding responsibility on the PIM DR proves a limitation in the protocol. In comparison, even though an OSPF DR, or an IS-IS DIS, handles additional duties while running the OSPF or IS-IS protocols, they are not required to be solely responsible for forwarding packets for the network. On the other hand, on a last hop LAN, only the PIM DR is asked to forward packets while the other routers handle only control traffic (and perhaps dropping packets due to RPF failures). The forwarding load of a last hop LAN is concentrated on a single router.

This leads to several issues. One of the issues is that the aggregated bandwidth will be limited to what R1 can handle towards this particular interface. These days, it is very common that the last hop LAN usually consists of switches that run IGMP/MLD or PIM snooping. This allows the forwarding of multicast packets to be restricted only to segments leading to receivers who have indicated their interest in multicast groups using either IGMP or MLD. The emergence of the switched Ethernet allows the aggregated bandwidth to exceed, some times by a large number, that of a single link. For example, let us modify Figure 1 and introduce an Ethernet switch in Figure 2.

```
                          ( core networks )
                            |     |     |
                            |     |     |
                           R1    R2     R3
                            |     |     |
                         +=gi0===gi1===gi2=+
                         +                 +
                         +      switch     +
                         +                 +
                         +=gi4===gi5===gi6=+
                            |     |     |
                           H1    H2     H3
```

Figure 2: Last Hop Network with Ethernet Switch

Let us assume that each individual link is a Gigabit Ethernet. Each router, R1, R2 and R3, and the switch have enough forwarding capacity that can handle hundreds of Gigabits of data.

Let us further assume that each of the hosts requests 500 mbps of data and different traffic is requested by each host. This represents a total 1.5 gbps of data, which is under what each switch or the combined uplink bandwidth across the routers can handle, even under failure of a single router.

On the other hand, the link between R1 and switch, via port gi0, can only handle a throughput of 1gbps. And if R1 is the only router, the PIM DR elected using the procedure defined by RFC 4601, at least 500 mbps worth of data will be lost because the only link that can be used to draw the traffic from the routers to the switch is via gi0. In other words, the entire network's throughput is limited by the single connection between the PIM DR and the switch (or the last hop LAN as in Figure 1).

The problem may also manifest itself in a different way. For example, R1 happens to forward 500 mbps worth of unicast data to H1, and at the same time, H2 and H3 each requests 300 mbps of different multicast data. Once again packet drop happens on R1 while in the mean time, there is sufficient forwarding capacity left on R2 and R3 and link capacity between the switch and R2/R3.

Another important issue is related to failover. If R1 is the only forwarder on the last hop network, in the event of a failure when R1 goes out of service, multicast forwarding for the entire network has to be rebuilt by the newly elected PIM DR. However, if there was a way that allowed multiple routers to forward to the network for different groups, failure to one of the routers would only lead to

disruption to a subset of the flows, therefore improving the overall resilience of the network.

In this document, we propose a modification to the PIM protocol that allows multiple of these routers, called Group Designated Router (GDR) to be selected so that the forwarding load can be distributed to and handled by a number of routers.

3. Applicability

The proposed change described in this draft applies to PIM last hop routers only.

It doesn't alter the behavior of a PIM DR on the first hop network. This is because the source tree is built using the IP address of the sender, not the IP address of the PIM DR that sends the registers towards the RP. The load balancing between first hop routers can be achieved naturally if an IGP provides equal cost multiple paths (which it usually does in practice). And distributing the load to do registering doesn't justify the additional complexity required to support it.

4. Functional Overview

In existing PIM DR election, when multiple last hop routers are connected to a multi-access network (for example, an Ethernet), one of them is selected to act as PIM DR. The PIM DR is responsible for sending Join/Prune messages to the RP or source. To elect the PIM DR, each PIM router on the network examines the received PIM Hello messages and compares its DR priority and IP address with those of its neighbors. The router with the highest DR priority is the PIM DR. If there are multiple such routers, IP address is used as the tie breaker, as described in [RFC4601].

In order to share forwarding load among last hop routers, besides the normal PIM DR election, the GDR is also elected on the last hop multi-access network. There is only one PIM DR on the multi-access network, but there might be multiple GDR candidates.

For each multicast group, a hash algorithm is used to select one of the routers to be the GDR. Hash Masks are defined for Source, Group and RP separately, in order to handle different PIM modes. The masks are announced in PIM Hello as a new Load Balancing Hash Mask TLV (LBM TLV). Last hop routers with this new TLV and with the same DR priority as the PIM DR are GDR candidates.

A simple hash algorithm based on the announced Source, Group or RP masks allow one GDR to be assigned to a corresponding multicast group, and that GDR is responsible for initiating the creation of multicast forwarding tree for the group.

4.1. GDR Candidates

GDR is the new concept introduced by this draft. To become a candidate GDR, a router must have the same DR priority as the DR. For example, if there are 4 routers on the LAN: R1, R2, R3 and R4. R1, R2 and R3 have the same DR priority while R4's DR priority is less preferred. In this example, only R1, R2 and R3 will be eligible for GDR election. R4 is not because R4 will not become a PIM DR unless all of R1, R2 and R3 go out of service.

Further assuming router R1 wins the PIM DR election. In its Hello packet, R1 will include the identity of R1, R2 and R3 (the GDR candidates) besides its own Load Balancing Hash Mask TLV. The order of the GDR candidates is converted to the ordinal number associated with each GDR candidate. For example, addresses advertised by R1 is R1, R2, R3, the ordinal number assigned to R1 is 0, to R2 is 1 and to R3 is 2.

4.2. Hash Mask

A Hash Mask is used to extract a number of bits from the corresponding IP address field (32 for v4, 128 for v6), and calculate a hash value. A hash value is used to select GDR from GDR Candidates advertised in order by PIM DR. For example, 0.255.0.0 defines a Hash Mask for an IPv4 address that masks the first, the third and the fourth octets.

There are three Hash Masks defined,

o RP Hash Mask
o Source Hash Mask
o Group Hash Mask

The Hash Masks must be configured on the PIM routers that can potentially become a PIM DR.

For ASM groups, a hash value is calculated using the following formula:

o hashvalue_RP = ((RP_address & RP_hashmask) >> N) % M

RP_address is the address of the RP defined for the group. N is the number of bits that are 0 from the right. M is the number of GDR

candidates as described above.

If RP_hashmask is 0, a hash value is also calculated using the group Hash Mask in a similar fashion

o hashvalue_Group = ((Group_address & Group_hashmask) >> N) % M

For SSM groups, a hash value is calculated using both the source and group Hash Mask

o hashvalue_SG = (((Source_address & Source_hashmask) >> N_S) ^
   ((Group_address & Group_hashmask) >> N_G)) % M

4.3. PIM Hello Options

When a non-DR PIM router that supports this draft sends a PIM Hello, it includes a new option, called "Load Balancing Hash Masks TLV (LBM TLV)". The LBM TLV consists of three Hash Masks as defined above.

Besides this new LBM TLV, the elected PIM DR router also includes a "Load Balancing GDR TLV (LBGDR TLV)" in its PIM Hello. The LBGDR TLV consists of the sorted addresses of all GDR candidates on the last hop network.

The elected PIM DR router uses LBM TLV to calculate its LBGDR TLV. The GDR candidates use LBM TLV and LBGDR TLV advertised by DR PIM router to calculate hash value.

5. Packet Format

5.1. PIM DR Load Balancing GDR (LBGDR) Hello TLV

```
     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 = TBD          |         Length                |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                          GDR Address(es)                      |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: GDR Hello TLV

Type: TBD

Length:

GDR Address (32/128 bits):   Address(es) of GDR candidates. All addresses must be in the same address family. The addresses are sorted from high to low. The order is used as the ordinal number, starting from 0, in hash value calculation.

This LBGDR TLV should only be advertised by the elected PIM DR router.

5.2. PIM DR Load Balancing Hash Masks (LBM) Hello TLV

```
        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 = TBD          |          Length               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                            Group Mask                         |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                            Source Mask                        |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                            RP Mask                            |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Hash Masks Hello TLV

Type: TBD.

Length:

Group Mask (32/128 bits):   Mask

Source Mask (32/128 bits):   Mask

RP Mask (32/128 bits):   Mask

All masks must be in the same address family, with the same length.

This LBM TLV should be advertised by last hop routers, which support this draft.

6. Protocol Specification

6.1. PIM DR Operation

The DR elect process is still the same as defined in [RFC4601]. A DR supports this draft advertises a new Hello Option LBGRD TLV to includes all GDR candidates. Moreover, same as non-DR routers, DR also advertises LBM TLV Hello Option to indicate its capability of supporting this draft.

LBGRD TLV is composed by sorting the addresses of all GDR candidates. LBM TLV on PIM DR contains value of masks from user configuration.

If a PIM DR receives a neighbor Hello with LBGRD TLV, the PIM DR should ignore the TLV.

If a PIM DR receives a neighbor Hello with LBM TLV, and the neighbor has the same DR priority as PIM DR itself, the PIM DR should consider the neighbor as a GDR candidate and insert the neighbor's address into the sorted list of LBGRD TLV.

6.2. PIM GDR Candidate Operation

When an IGMP join is received, without this proposal, router R1 (the PIM DR) will handle the join and potentially run into the issues described earlier. Using this proposal, a simple algorithm is used to determine which router is going to be responsible for building forwarding trees on behalf of the host.

The algorithm works as follows, assuming the router in question is X and a GDR Candidate:

- If the group is ASM, and if the RP Hash Mask announced by the PIM DR is not 0.0.0.0, calculate the value of hashvalue_RP. If hashvalue_RP is the same as the ordinal number assigned implicitly to X by PIM DR, X becomes the GDR.
- If the group is ASM and if the RP Hash Mask announced by the PIM DR is 0, obtain the value of hashvalue_Group. And compare that to the ordinal value of X assigned by the PIM DR to decide if X is the GDR
- If the group is SSM, then use hashvalue_SG to determine if X is the GDR.

If X is the GDR for the group, X will be responsible for building the forwarding tree.

A router that supports this draft advertises LBM TLV in its Hello, even the router may not be a GDR candidate.

A GDR candidate may receive a LBM TLV from PIM DR router, with a

different Hash Masks as advertised in its own Hello LBM TLV. The GDR candidate must use the Hash Masks advertised by the PIM DR Hello to calculate the hash value.

A GDR candidate may receive a LBGDR TLV from a non-DR PIM router. The GDR candidate must ignore such LBGDR TLV.

A GDR candidate may receive Hello from the elected PIM DR, and the PIM DR doesn't support this draft. The GDR election described by this draft will not take place, that is only the PIM DR joins the multicast tree.

6.3. PIM Assert Modification

When routers restart, GDR may change for a specific group, which might cause packet drops.

For example, if there are two streams G1 and G2, and R1 is the GDR for G1 and R2 is the GDR for G2. When R3 comes up online, it is possible that R2 becomes GDR for G1 and R3 becomes GDR for G2, and rebuilding of the forwarding trees for G1 and G2 will lead to potential packet loss.

This is not a typical deployment scenario but it still might happen. Here we describe a mechanism to minimize the impact.

When the role of GDR changes as above, instead of immediately stopping forwarding, R1 and R2 continue forwarding to G1 and G2 respectively, while in the same time, R2 and R3 build forwarding trees for G1 and G2 respectively. This will lead to PIM Asserts.

The same tie breakers are used to select an Assert winner with one modification. That is, instead of comparing IP addresses as the last resort, a router considers whether the sender of an Assert is a GDR. In this example, R1 will let R2 be the assert winner for G1, and R2 will do the same for R3 for G2. This will cause some duplicates in the network while minimizing packet loss.

If a router on the LAN doesn't support this draft, the Assert modification described above will not take place, that is only the IP address of an Assert sender is used as the tie breaker. Fore example, if R4, with preferred IP address, doesn't understand GDR and sends Assert for G1, R2, the GDR for G1, will grant R4 as the Assert winner, clear OIF on R2.

7. IANA Considerations

Two new PIM Hello Option Types are required to assign to the DR Load Balancing messages. According to [HELLO-OPT], this document recommends 32(0x20) as the new "PIM DR Load Balancing GDR Hello Option", and 33(0x21) as the new "PIM DR Load Balancing Hash Masks Hello Option" .

8. Security Considerations

Security of the PIM DR Load Balancing Hello message is only guaranteed by the security of PIM Hello packet, so the security considerations for PIM Hello packets as described in PIM-SM [RFC4601] apply here.

9. Acknowledgement

The authors would like to thank Steve Simlo, Taki Millonis for helping with the original idea, ??? for their review comments.

10. References

10.1. Normative Reference

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

10.2. Informative References

[RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, January 2005.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

[HELLO-OPT] IANA, "PIM Hello Options", PIM-HELLO-OPTIONS per RFC4601 http://www.iana.org/assignments/pim-hello-options, March 2007.

Authors' Addresses

Yiqun Cai
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: ycai@cisco.com

Sri Vallepalli
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: svallepa@cisco.com

Heidi Ou
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: hou@cisco.com

Andy Green
British Telecom
Adastral Park
Ipswich IP5 2RE
United Kingdom

Email: andy.da.green@bt.com