Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 6, 2020

Y. Cai
H. Ou
Alibaba Group
S. Vallepalli
M. Mishra
S. Venaas
Cisco Systems, Inc.
A. Green
British Telecom

January 3, 2020

PIM Designated Router Load Balancing
draft-ietf-pim-drlb-15

Abstract

On a multi-access network, one of the PIM-SM (PIM Sparse Mode) routers is elected as a Designated Router. One of the responsibilities of the Designated Router is to track local multicast listeners and forward data to these listeners if the group is operating in PIM-SM. This document specifies a modification to the PIM-SM protocol that allows more than one of the PIM-SM routers to take on this responsibility so that the forwarding load can be distributed among multiple routers.

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, 2020.

Table of Contents

1. Introduction

On a multi-access LAN, such as an Ethernet, with one or more PIM-SM (PIM Sparse Mode) [RFC7761] routers, one of the PIM-SM routers is elected as a Designated Router (DR). 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) if the group is operating in PIM-SM. Also, the PIM DR is responsible for tracking local multicast listeners and forwarding to these listeners if the group is operating in PIM-SM.

Consider the following LAN in Figure 1:

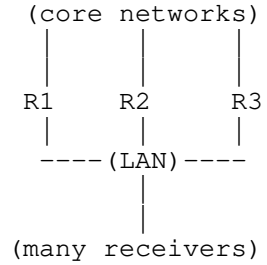

Figure 1: LAN with receivers

Assume R1 is elected as the DR. According to the PIM-SM protocol, R1 will be responsible for forwarding traffic to that LAN on behalf of all local members. In addition to keeping track of membership reports, R1 is also responsible for initiating the creation of source and/or shared trees towards the senders or the RPs. The membership reports would be IGMP or MLD messages. This applies to any versions of the IGMP and MLD protocols. The most recent versions are IGMPv3 [RFC3376] and MLDv2 [RFC3810].

Having a single router acting as DR and being responsible for data plane forwarding leads to several issues. One of the issues is that the aggregated bandwidth will be limited to what R1 can handle with regards to capacity of incoming links, the interface on the LAN, and total forwarding capacity. It is very common that a LAN consists of switches that run IGMP/MLD or PIM snooping [RFC4541]. This allows the forwarding of multicast packets to be restricted only to segments leading to receivers that have indicated their interest in multicast groups using either IGMP or MLD. The emergence of the switched Ethernet allows the aggregated bandwidth to exceed, sometimes by a large number, that of a single link. For example, let us modify Figure 1 and introduce an Ethernet switch in Figure 2.

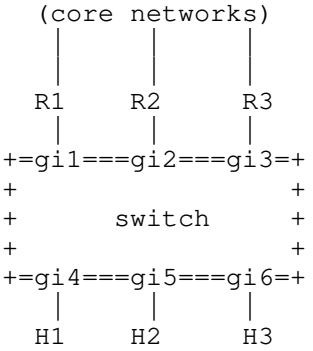

Figure 2: LAN with Ethernet Switch

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

Let us further assume that each of the hosts requests 500 Mbps of unique multicast data. This totals to 1.5 Gbps of data, which is less than 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 gi1, can only handle a throughput of 1Gbps. And if R1 is the only DR (the PIM DR elected using the procedure defined by [RFC7761]) 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 gi1. In other words, the entire network's throughput is limited by the single connection between the PIM DR and the switch (or LAN as in Figure 1).

Another important issue is related to failover. If R1 is the only forwarder on a shared LAN, when R1 goes out of service, multicast forwarding for the entire LAN has to be rebuilt by the newly elected PIM DR. However, if there were a way that allowed multiple routers to forward to the LAN for different groups, failure of one of the routers would only lead to disruption to a subset of the flows, therefore improving the overall resilience of the network.

This document specifies a modification to the PIM-SM protocol that allows more than one of these routers, called Group Designated

Routers (GDR) to be selected so that the forwarding load can be distributed among a number of routers.

2. Terminology

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.

With respect to PIM-SM, this document follows the terminology that has been defined in [RFC7761].

This document also introduces the following new acronyms:

- GDR: Group Designated Router. For each multicast flow, either a (*,G) for Any-Source Multicast (ASM), or an (S,G) for Source-Specific Multicast (SSM) [RFC4607], a Hash Algorithm (described below) is used to select one of the routers as a GDR. The GDR is responsible for initiating the forwarding tree building process for the corresponding multicast flow.

- GDR Candidate: a router that has the potential to become a GDR. There might be multiple GDR Candidates on a LAN, but only one can become the GDR for a specific multicast flow.

3. Applicability

The extension specified in this document applies to PIM-SM routers acting as last hop routers (there are directly connected receivers). It does not alter the behavior of a PIM DR, or any other routers, on the first hop network (directly connected sources). 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 PIM 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). Also distributing the load to do source registration does not justify the additional complexity required to support it.

4. Functional Overview

In the PIM DR election as defined in [RFC7761], when multiple routers are connected to a multi-access LAN (for example, an Ethernet), one of them is elected to act as PIM DR. The PIM DR is responsible for sending local Join/Prune messages towards the RP or source. In order to elect the PIM DR, each PIM router on the LAN examines the received

PIM Hello messages and compares its own 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, their IP addresses are used as the tie-breaker, as described in [RFC7761].

In order to share forwarding load among last hop routers, besides the normal PIM DR election, one or more GDRs are elected on the multi-access LAN. There is only one PIM DR on the multi-access LAN, but there might be multiple GDR Candidates.

For each multicast flow, that is, (*,G) for ASM and (S,G) for SSM, a Hash Algorithm [Section 5.1] is used to select one of the routers to be the GDR. The new DR Load Balancing Capability (DRLB-Cap) PIM Hello Option is used to announce the Capability as well as the Hash Algorithm type. Routers with the new DRLB-Cap Option advertised in their PIM Hello, using the same GDR election Hash Algorithm and the same DR priority as the PIM DR, are considered as GDR Candidates.

Hash Masks are defined for Source, Group and RP separately, in order to handle PIM ASM/SSM. The masks, as well as a sorted list of GDR Candidate Addresses, are announced by the DR in a new DR Load Balancing List (DRLB-List) PIM Hello Option.

A Hash Algorithm based on the announced Source, Group, or RP masks allows one GDR to be assigned to a corresponding multicast state. That GDR is responsible for initiating the creation of the multicast forwarding tree for multicast traffic.

4.1. GDR Candidates

GDR is the new concept introduced by this specification. GDR Candidates are routers eligible for GDR election on the LAN. To become a GDR Candidate, a router must have the same DR priority and run the same GDR election Hash Algorithm as the DR on the LAN.

For example, assume there are 4 routers on the LAN: R1, R2, R3 and R4, each announcing a DRLB-Cap option. R1, R2 and R3 have the same DR priority while R4's DR priority is less preferred. In this example, R4 will not be eligible for GDR election, because R4 will not become a PIM DR unless all of R1, R2 and R3 go out of service.

Furthermore, assume router R1 wins the PIM DR election, R1 and R2 advertise the same Hash Algorithm for GDR election, while R3 advertises a different one. In this case, only R1 and R2 will be eligible for GDR election, while R3 will not.

As a DR, R1 will include its own Load Balancing Hash Masks and the identity of R1 and R2 (the GDR Candidates) in its DRLB-List Hello Option.

5. Protocol Specification

5.1. Hash Mask and Hash Algorithm

A Hash Mask is used to extract a number of bits from the corresponding IP address field (32 for IPv4, 128 for IPv6) and calculate a hash value. A hash value is used to select a GDR from GDR Candidates advertised by the PIM DR. Hash masks allow for certain flows to always be forwarded by the same GDR, by ignoring certain bits in the hash value calculation, so that the hash values are the same. For example, 0.0.255.0 defines a Hash Mask for an IPv4 address that masks the first, the second, and the fourth octets, which means that only the third octet will influence the hash value computed. Note that the masks need not be a contiguous set of bits. E.g, for IPv4, 15.15.15.15 would be a valid mask.

In the text below, a hash mask is in some places said to be zero. A hash mask is zero if no bits are set. That is, 0.0.0.0 for IPv4 and :: for IPv6. Also, a hash mask is said to be an all-bits-set mask if it is 255.255.255.255 for IPv4 or ffff:ffff:ffff:ffff:ffff:ffff:ffff:ffff for IPv6.

There are three Hash Masks defined:

- RP Hash Mask
- Source Hash Mask
- Group Hash Mask

The hash masks need to be configured on the PIM routers that can potentially become a PIM DR, unless the implementation provides default hash mask values. An implementation SHOULD have default hash mask values as follows. The default RP Hash Mask SHOULD be zero (no bits set). The default Source and Group Hash Masks SHOULD both be all-bits-set masks. These default values are likely acceptable for most deployments, and simplify configuration. There is only a need to use other masks if one needs to ensure that certain flows are forwarded by the same GDR.

The DRLB-List Hello Option contains a list of GDR Candidates. The first one listed has ordinal number 0, the second listed ordinal number 1, and the last one has ordinal number N - 1 if there are N candidates listed. The hash value computed will be the ordinal

number of the GDR Candidate that is acting as GDR for the flow in question.

The input to be hashed is determined as follows:

- If the group is in ASM mode and the RP Hash Mask announced by the PIM DR is not zero (at least one bit is set), calculate the value of hashvalue_RP [Section 5.2] to determine the GDR.

- If the group is in ASM mode and the RP Hash Mask announced by the PIM DR is zero (no bits are set), obtain the value of hashvalue_Group [Section 5.2] to determine the GDR.

- If the group is in SSM mode, use hashvalue_SG [Section 5.2] to determine the GDR.

A simple Modulo Hash Algorithm is defined in this document. However, to allow another Hash Algorithms to be used, a 1-octet "Hash Algorithm" field is included in the DRLB-Cap Hello Option to specify the Hash Algorithm used by the router.

If different Hash Algorithms are advertised among the routers on a LAN, only the routers advertising the same Hash Algorithm as the DR (as well as having the same DR priority as the DR) are eligible for GDR election.

5.2. Modulo Hash Algorithm

As part of computing the hash, the notation LSZC(hash_mask) is used to denote the number of zeroes counted from the least significant bit of a Hash Mask hash_mask. As an example, LSZC(255.255.128) is 7 and also LSZC(ffff:8000::) is 111. If all bits are set, LSZC will be 0. If the mask is zero, then LSZC will be 32 for IPv4, and 128 for IPv6.

The number of GDR Candidates is denoted as GDRC.

The idea behind the Modulo Hash Algorithm is in simple terms that the corresponding mask is applied to a value, then the result is shifted right LSZC(mask) bits so that the least significant bits that were masked out are not considered. Then this result is masked by 0xffffffff, keeping only the last 32 bits of the result (this only makes a difference for IPv6). Finally, the hash value is this result modulo the number of GDR Candidates (GDRC).

The Modulo Hash Algorithm for computing the values hashvalue_RP, hashvalue_Group and hashvalue_SG is defined as follows.

hashvalue_RP is calculated as:

(((RP_address & RP_mask) >> LSZC(RP_mask)) & 0xffffffff) % GDRC

RP_address is the address of the RP defined for the group and RP_mask is the RP Hash Mask.

hashvalue_Group is calculated as:

(((Group_address & Group_mask) >> LSZC(Group_mask)) & 0xffffffff) % GDRC

Group_address is the group address and Group_mask is the Group Hash Mask.

hashvalue_SG is calculated as:

((((Source_address & Source_mask) >> LSZC(Source_mask)) & 0xffffffff) ^ (((Group_address & Group_mask) >> LSZC(Group_mask)) & 0xffffffff)) % GDRC

Group_address is the group address and Group_mask is the Group Hash Mask.

5.2.1. Modulo Hash Algorithm Examples

To help illustrate the algorithm, consider this example. Router X with IPv4 address 203.0.113.1 receives a DRLB-List Hello Option from the DR, which announces RP Hash Mask 0.0.255.0 and a list of GDR Candidates, sorted by IP addresses from high to low: 203.0.113.3, 203.0.113.2 and 203.0.113.1. The ordinal number assigned to those addresses would be:

0 for 203.0.113.3; 1 for 203.0.113.2; 2 for 203.0.113.1 (Router X).

Assume there are 2 RPs: RP1 192.0.2.1 for Group1 and RP2 198.51.100.2 for Group2. Following the modulo Hash Algorithm:

LSZC(0.0.255.0) is 8 and GDRC is 3. The hashvalue_RP for Group1 with RP RP1 is:

(((192.0.2.1 & 0.0.255.0) >> 8) & 0xffffffff % 3) = 2 % 3 = 2

which matches the ordinal number assigned to Router X. Router X will be the GDR for Group1.

The hashvalue_RP for Group2 with RP RP2 is:

(((198.51.100.2 & 0.0.255.0) >> 8) & 0xffffffff % 3) = 100 % 3 = 1

which is different from the ordinal number of Router X (2). Hence, Router X will not be GDR for Group2.

For IPv6 consider this example, similar to the above. Router X with IPv6 address fe80::1 receives a DRLB-List Hello Option from the DR, which announces RP Hash Mask ::ffff:ffff:ffff:0 and a list of GDR Candidates, sorted by IP addresses from high to low: fe80::3, fe80::2 and fe80::1. The ordinal number assigned to those addresses would be:

0 for fe80::3; 1 for fe80::2; 2 for fe80::1 (Router X).

Assume there are 2 RPs: RP1 2001:db8::1:0:5678:1 for Group1 and RP2 2001:db8::1:0:1234:2 for Group2. Following the modulo Hash Algorithm:

LSZC(::ffff:ffff:ffff:0) is 16 and GDRC is 3. The hashvalue_RP for Group1 with RP RP1 is:

(((2001:db8::1:0:5678:1 & ::ffff:ffff:ffff:0) >> 16) & 0xffffffff % 3) = ((::1:0:5678:0 >> 16) & 0xffffffff % 3) = (::1:0:5678 & 0xffffffff % 3) = ::5678 % 3 = 2

which matches the ordinal number assigned to Router X. Router X will be the GDR for Group1.

The hashvalue_RP for Group2 with RP RP2 is:

(((2001:db8::1:0:1234:1 & ::ffff:ffff:ffff:0) >> 16) & 0xffffffff % 3) = ((::1:0:1234:0 >> 16) & 0xffffffff % 3) = (::1:0:1234 & 0xffffffff % 3) = ::1234 % 3 = 1

which is different from the ordinal number of Router X (2). Hence, Router X will not be GDR for Group2.

5.2.2. Limitations

The Modulo Hash Algorithm has poor failover characteristics when a shared LAN has more than two GDRs. In the case of more than two GDRs on a LAN, when one GDR fails, all of the groups may be reassigned to a different GDR, even if they were not assigned to the failed GDR. However, many deployments use only two routers on a shared LAN for redundancy purposes. Future work may define new Hash Algorithms where only groups assigned to the failed GDR get reassigned.

The Modulo Hash Algorithm will use at most 32 consecutive bits of the input addresses for its computation. Exactly which bits are used of the source, group or RP addresses, depend on the respective masks.

This limitation may be an issue for IPv6 deployments, since not all bits of the IPv6 addresses are considered. If this causes operational issues, a new hash algorithm would need to be defined.

5.3. PIM Hello Options

PIM routers include a new option, called "Load Balancing Capability (DRLB-Cap)" in their PIM Hello messages.

Besides this DRLB-Cap Hello Option, the elected PIM DR also includes a new "DR Load Balancing List (DRLB-List) Hello Option". The DRLB-List Hello Option consists of three Hash Masks as defined above and also a list of GDR Candidate addresses on the LAN. It is recommended that the GDR Candidate addresses are sorted in descending order. This ensures that when using algorithms such as the Modulo algorithm in this document, that it is predictable which GDR is responsible for which groups, regardless of the order the DR learned about the candidates.

5.3.1. PIM DR Load Balancing Capability (DRLB-Cap) Hello Option

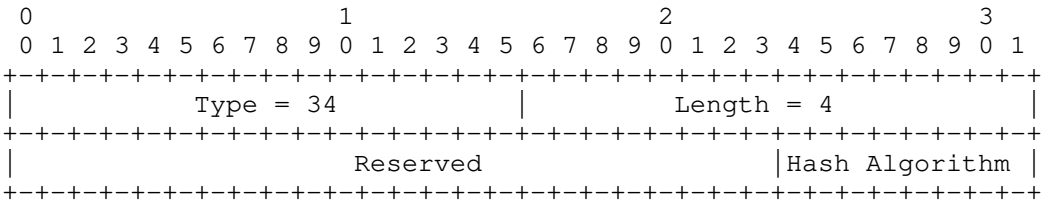

Figure 3: PIM DR Load Balancing Capability Hello Option

Type: 34

Length: 4

Reserved: Transmitted as zero, ignored on receipt.

Hash Algorithm: Hash Algorithm type. A value listed in the IANA Designated Router Load Balancing Hash Algorithms registry. 0 is used for the Modulo algorithm defined in this document.

This DRLB-Cap Hello Option MUST be advertised by routers on all interfaces where DR Load Balancing is enabled. Note that the option is included at most once.

5.3.2. PIM DR Load Balancing List (DRLB-List) Hello Option

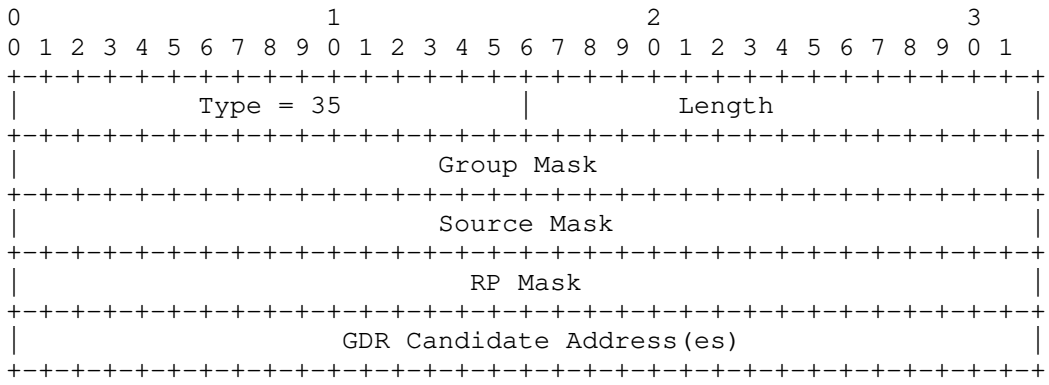

Figure 4: PIM DR Load Balancing List Hello Option

Type: 35

Length: (3 + n) x (4 or 16) bytes, where n is the number of GDR candidates.

Group Mask (32/128 bits): Mask applied to group addresses as part of hash computation.

Source Mask (32/128 bits): Mask applied to source addresses as part of hash computation.

RP Mask (32/128 bits): Mask applied to RP addresses as part of hash computation.

All masks MUST have the same number of bits as the IP source address in the PIM Hello IP header.

GDR Candidate Address(es) (32/128 bits): List of GDR Candidate(s)

All addresses MUST be in the same address family as the PIM Hello IP header. It is recommended that the addresses are sorted in descending order.

If the "Interface ID" option, as specified in [RFC6395], is present in a GDR Candidate's PIM Hello message, and the "Router Identifier" portion is non-zero:

+ For IPv4, the "GDR Candidate Address" will be set directly to the "Router Identifier".

+ For IPv6, the "GDR Candidate Address" will be 96 bits of zeroes followed by the 32 bit Router Identifier.

If the "Interface ID" option is not present in a GDR Candidate' PIM Hello message, or if the "Interface ID" option is present but the "Router Identifier" field is zero, the "GDR Candidate Address" will be the IPv4 or IPv6 source address of the PIM Hello message.

This DRLB-List Hello Option MUST only be advertised by the elected PIM DR. It MUST be ignored if received from a non-DR. The option MUST also be ignored if the hash masks are not the correct number of bits, or GDR Candidate addresses are in the wrong address family.

5.4. PIM DR Operation

The DR election process is still the same as defined in [RFC7761]. The DR advertises the new DRLB-List Hello Option, which contains mask values from user configuration (or default values), followed by a list of GDR Candidate Addresses. Note that if a router included the "Interface ID" option in the hello message, and the Router ID is non-zero, the Router ID will be used to form the GDR Candidate address of the router, as discussed in the previous section. It is recommended that the list be sorted, from the highest value to the lowest value. The reason for sorting the list is to make the behavior deterministic, regardless of the order in which the DR learns of new candidates. Note that, as for non-DR routers, the DR also advertises the DRLB-Cap Hello Option to indicate its ability to support the new functionality and the type of GDR election Hash Algorithm it uses.

If a PIM DR receives a neighbor DRLB-Cap Hello Option, which contains the same Hash Algorithm as the DR, and the neighbor has the same DR priority as the DR, PIM DR SHOULD consider the neighbor as a GDR Candidate and insert the GDR Candidate' Address into the list of the DRLB-List Option. However, the DR may have policies limiting which GDR Candidates, or the number of GDR Candidates to include. Likewise, the DR SHOULD include itself in the list of GDR Candidates, but it is permissible not to do so, if for instance there is some policy restricting the candidate set.

If a PIM neighbor included in the list expires, stops announcing the DRLB-Cap Hello Option, changes DR priority, changes Hash Algorithm or otherwise becomes ineligible as a candidate, the DR SHOULD

immediately send a triggered hello with a new list in the DRLB-List option, excluding the neighbor.

If a new router becomes eligible as a candidate, there is no urgency in sending out an updated list. An updated list SHOULD be included in the next hello.

5.5. PIM GDR Candidate Operation

When an IGMP/MLD report is received, a Hash Algorithm is used by the GDR Candidates to determine which router is going to be responsible for building forwarding trees on behalf of the host.

The router MUST include the DRLB-Cap Hello Option in all PIM Hello messages sent on the interface. Note that the presence of the DRLB-Cap Option in the PIM Hello does not guarantee that the router will be considered as a GDR candidate. Once the DR election is done, the DRLB-List Hello Option is received from the current PIM DR containing a list of the selected GDRs Candidates.

A router only acts as a GDR Candidate if it is included in the GDR Candidate list of the DRLB-List Hello Option. See next section for details.

5.6. DRLB-List Hello Option Processing

This section discusses processing of the DRLB-List Hello Option, including the case where it was received in the previous hello, but not in the current hello. All routers MUST ignore the DRLB-List Hello Option if it is received from a PIM router which is not the DR. The option MUST only be processed by routers that are announcing the DRLB-Cap Option, and only if the Hash Algorithm announced by the DR is the same as the local announcement. All GDR Candidates MUST use the Hash Masks advertised in the Option, even if they differ from those the candidate was configured with. The DR MUST also process its own DRLB-List Hello Option.

A router stores the latest option contents that was announced, if any, and deletes the previous contents. The router MUST also compare the new contents with any previous contents, and if there are any changes, continue processing as below. Note that if the option does not pass the above checks, the below processing MUST be done as if the option was not announced.

If the contents of the DRLB-List Option, the masks or the candidate list, differs from the previously saved copy, it is received for the first time, or it is no longer being received or accepted, the option MUST be processed as below.

1. If the local router is included in the GDR Candidate Address(es) field (it will look for its own address, or its Router ID if it announces a non-zero Router ID), for each of the groups, or source and group pairs if the group is in SSM mode, with local receiver interest, the router MUST run the Hash Algorithm to determine which of them it is the GDR for.

   If there is no change in the GDR status, then no further action is required.

   If the router becomes the new GDR, then a multicast forwarding tree MUST be built [RFC7761].

   If the router is no longer the GDR, then it uses an Assert as explained in [Section 5.7].

2. If the local router is not included in the GDR Candidate Address(es) field, or if the DRLB-List Hello Option is no longer included in the DR's Hello, or if the DR's Neighbor Liveness Timer expires [RFC7761], for each of the groups, or source and group pairs if the group is in SSM mode, with local receiver interest, for which the router is the GDR, it uses an Assert as explained in [Section 5.7].

5.7. PIM Assert Modification

GDR changes may occur due to configuration change, due to GDR candidates going down, and also new routers coming up and becoming GDR candidates. This may occur while flows are being forwarded. If the GDR for an active flow changes, there is likely to be some disruption, such as packet loss or duplicates. By using asserts, packet loss is minimized, while allowing a small amount of duplicates.

When a router stops acting as the GDR for a group, or source and group pair if SSM, it MUST set the Assert metric preference to maximum (0x7fffffff) and the Assert metric to one less than maximum (0xfffffffe). That is, whenever it sends or receives an Assert for the group, it must use these values as the metric preference and metric rather than the values provided by the unicast routing protocol.

The rest of this section is just for illustration purposes and not part of the protocol definition.

To illustrate the behavior when there is a GDR change, consider the following scenario where there are two flows G1 and G2. R1 is the GDR for G1, and R2 is the GDR for G2. When R3 comes up, it is

possible that R3 becomes GDR for both G1 and G2, hence R3 starts to build the forwarding tree for G1 and G2. If R1 and R2 stop forwarding before R3 completes the process, packet loss might occur. On the other hand, if R1 and R2 continue forwarding while R3 is building the forwarding trees, duplicates might occur.

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

For G1, using the functionality described in this document, R1 and R3 determine the new GDR, which is R3. With the modified Assert behavior, R1 sets its Assert metric to the near maximum value discussed above. That will make R3, which has normal metric in its Assert as the Assert winner.

5.8. Backward Compatibility

In the case of a hybrid Ethernet shared LAN (where some PIM routers support the functionality defined in this document, and some do not);

- If the DR does not support the new functionality, then there will be no load-balancing.
- If non-DR routers do not support the new functionality, they will not be considered as Candidate GDRs and it will not take part in load-balancing. Load-balancing may still happen on the link.

6. Operational Considerations

An administrator needs to consider what the total bandwidth requirements are and find a set of routers that together has enough available capacity, while making sure that each of the routers can handle its part, assuming that the traffic is distributed roughly equally among the routers. Ideally, one should also have enough bandwidth to handle the case where at least one router fails. All routers should have reachability to the sources, and RPs if applicable, that is not via the LAN.

Care must be taken when choosing what hash masks to configure. One would typically configure the same masks on all the routers, so that they are the same, regardless of which router is elected as DR. The default masks are likely suitable for most deployment. The RP Hash Mask must be configured (the default is no bits set) if one wishes to hash based on the RP address rather than the group address for ASM. The default masks will use the entire group addresses, and source addresses if SSM, as part of the hash. An administrator may set

other masks that masks out part of the addresses to ensure that certain flows always get hashed to the same router. How this is achieved depends on how the group addresses are allocated.

Only the routers announcing the same Hash Algorithm as the DR would be considered as GDR candidates. Network administrators need to make sure that the desired set of routers announce the same algorithm. Migration between different algorithms is not considered in this document.

7. IANA Considerations

IANA has temporarily assigned type 34 for the PIM DR Load Balancing Capability (DRLB-Cap) Hello Option, and type 35 for the PIM DR Load Balancing List (DRLB-List) Hello Option in the PIM-Hello Options registry. IANA is requested to make these assignments permanent when this document is published as an RFC. Note that the option names have changed slightly since the temporary assignments were made. Also, the length of option 34 is always 4, the registry currently says it is variable.

This document requests IANA to create a registry called "Designated Router Load Balancing Hash Algorithms" in the "Protocol Independent Multicast (PIM)" branch of the registry tree. The registry lists Hash Algorithms for use by PIM Designated Router Load Balancing.

7.1. Initial registry

The initial content of the registry should be as follows.

Type	Name	Reference
0	Modulo	This document
1-255	Unassigned	

7.2. Assignment of new Hash Algorithms

Assignment of new Hash Algorithms is done according to the "IETF Review" model, see [RFC8126].

8. Security Considerations

Security of the new DR Load Balancing PIM Hello Options is only guaranteed by the security of PIM Hello messages, so the security

considerations for PIM Hello messages as described in PIM-SM [RFC7761] apply here.

If the DR is subverted it could omit or add certain GDRs or announce an unsupported algorithm. If another router is subverted, it could be made DR and cause similar issues. While these issues are specific to this specification, they are not that different from existing attacks such as subverting a DR and lowering the DR priority, causing a different router to become the DR.

If for any reason, the DR includes a GDR in the announced list which announces a different algorithm from what the DR announces, the GDR is required to ignore the announcement, and there will be no router acting as the DR for the flows that hash to that GDR.

If a GDR is subverted, it could potentially be made to stop forwarding all the traffic it is expected to forward. This is also similar today to if a DR is subverted.

An administrator may be able to achieve the desired load-balancing of known flows, but an attacker may send a single high rate flow which is served by a single GDR, or send multiple flows that are expected to be hashed to the same GDR.

9. Acknowledgement

The authors would like to thank Steve Simlo and Taki Millonis for helping with the original idea; Alia Atlas, Bill Atwood, Joe Clarke, Alissa Cooper, Jake Holland, Bharat Joshi, Anish Kachinthaya, Anvitha Kachinthaya, Benjamin Kaduk, Mirja Kuhlewind, Barry Leiba, Ben Niven-Jenkins, Alvaro Retana, Adam Roach, Michael Scharf, Eric Vyncke and Carl Wallace for reviews and comments; and Toerless Eckert and Rishabh Parekh for helpful conversation on the document.

10. References

10.1. 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>.

[RFC6395] Gulrajani, S. and S. Venaas, "An Interface Identifier (ID) Hello Option for PIM", RFC 6395, DOI 10.17487/RFC6395, October 2011, <https://www.rfc-editor.org/info/rfc6395>.

[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>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[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>.

10.2. Informative References

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <https://www.rfc-editor.org/info/rfc3376>.

[RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <https://www.rfc-editor.org/info/rfc3810>.

[RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, DOI 10.17487/RFC4541, May 2006, <https://www.rfc-editor.org/info/rfc4541>.

[RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, DOI 10.17487/RFC4607, August 2006, <https://www.rfc-editor.org/info/rfc4607>.

Authors' Addresses

Yiqun Cai
Alibaba Group

Email: yiqun.cai@alibaba-inc.com

Heidi Ou
Alibaba Group

Email: heidi.ou@alibaba-inc.com

Sri Vallepalli
Cisco Systems, Inc.
3625 Cisco Way
San Jose CA 95134
USA

Email: svallepa@cisco.com

Mankamana Mishra
Cisco Systems, Inc.
821 Alder Drive,
Milpitas CA 95035
USA

Email: mankamis@cisco.com

Stig Venaas
Cisco Systems, Inc.
Tasman Drive
San Jose CA 95134
USA

Email: stig@cisco.com

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

Email: andy.da.green@bt.com