Abstract

On a multi-access network, one of the PIM-SM 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.
1. Introduction

On a multi-access LAN, such as an Ethernet, with one or more PIM-SM 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:

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

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
any 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 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, 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.
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 gi0, 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 gi0. 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 was 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 when they act 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 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). Also distributing the load to do registering 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, the GDR is also 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 is used to select one of the routers to be the GDR. A new DR Load Balancing Capability (DRLB-Cap) PIM Hello Option, which contains hash algorithm type, is announced by routers on interfaces where this specification is enabled. 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. And 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 run the same hash algorithm for GDR election, while R3 runs 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 PIM DR. For example, 0.0.255.0 defines a Hash Mask for an IPv4 address that masks the first, the second, and the fourth octets. Hash masks allow for certain flows to always be forwarded by the same GDR, since the hash values are the same. For instance, the mask 0.0.255.0 means that only the third octet will be considered when hashing.

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.

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.
o 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.

o 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.

o 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 outers 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 0xFFFF, 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\textsubscript{Group} is calculated as:

$$\text{(((Group\_address \& Group\_mask) >> LSZC(\text{Group\_mask})) \& 0xFFFF) \% GDRC}$$

Group\_address is the group address and Group\_mask is the Group Hash Mask.

hashvalue\textsubscript{SG} is calculated as:

$$\text{(((Source\_address \& Source\_mask) >> LSZC(Source\_mask)) \& 0xFFFF) \^ (((Group\_address \& Group\_mask) >> LSZC(Group\_mask)) \& 0xFFFF)) \% GDRC}$$

Group\_address is the group address and Group\_mask is the Group Hash Mask.

5.2.1. Modulo Hash Algorithm Example

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:

$$\text{(((192.0.2.1 \& 0.0.255.0) >> 8) \& 0xFFFF \% 3) = 2 \% 3 = 2}$$

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:

$$\text{(((198.51.100.2 \& 0.0.255.0) >> 8) \& 0xFFFF \% 3) = 100 \% 3 = 1}$$

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.

5.3. PIM Hello Options

When a PIM router sends a PIM Hello on an interface with this specification enabled, it includes a new option, called "Load Balancing Capability (DRLB-Cap)".

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 sorted list of GDR Candidate addresses on the LAN.

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

```
 0                   1                   2                   3
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Type = 34           |         Length = 4            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Reserved                  |Hash Algorithm |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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. 0 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.
## 5.3.2. PIM DR Load Balancing List (DRLB-List) Hello Option

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Type = 35           |         Length                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Group Mask                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Source Mask                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            RP Mask                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    GDR Candidate Address(es)                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

**Figure 4: PIM DR Load Balancing List Hello Option**

**Type:** 35

**Length:** $(3 + n) \times (4 \text{ or } 16)$, 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 Address (32/128 bits):** Address(es) 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 ID" portion is non-zero:
For IPv4, the "GDR Candidate Address" will be set directly to the "Router ID".

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

If the "Interface ID" option is not present in a GDR Candidate’s PIM Hello message, or if the "Interface ID" option is present but the "Router ID" 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.

5.4. PIM DR Operation

The DR election process is still the same as defined in [RFC7761]. A DR that has this specification enabled on an interface 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. It is RECOMMENDED that the list is 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 the DR learns of new candidates. Note that same as non-DR routers, the DR also advertises DRLB-Cap Hello Option to indicate its capability of supporting this specification and the type of its GDR election hash algorithm.

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. The DR would normally include itself in the list of GDR Candidates.

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, without this specification, only the PIM DR will handle the join and potentially run into the issues described earlier. Using this specification, 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.

If this specification is enabled on an interface, the router MUST include the DRLB-Cap Hello Option in all PIM Hello messages sent on that interface. Note that the presence of the DRLB-Cap Option in PIM Hello does not guarantee that this router would be considered as a GDR candidate. Once DR election is done, the DRLB-List Hello Option would be received from the current PIM DR on the link which would contain a list of GDRs Candidates selected by the PIM DR.

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. 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. Also, the algorithm announced in the DRLB-Cap Option, MUST be the same as what was announced by the DR. All GDR Candidates MUST use the Hash Masks advertised in the Option, even if they differ from those the candidate was configured with.

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 router was not included in the previous GDR list, or there was no previous GDR list, but it is included in the new GDR list, the router MUST for each of the groups, or source and group pairs if the group is in SSM mode, with local receiver interest, run the hash algorithm to determine which of them it is the GDR for.
If it is not the GDR for a group, or source and group pair if SSM, no processing is required.

If it is hashed as the GDR, it needs to build a multicast forwarding tree.

2. If the router was included in the previous GDR list, and still is included in the new GDR list: The router MUST for each of the groups, or source and group pairs if the group is in SSM mode, with local receiver interest, run the hash algorithm to determine which of them it is the GDR for.

   If it was the GDR for a group, or source and group pair if SSM, and the new hash result chose it as the GDR, then no processing is required.

   If it was the GDR for a group, or source and group pair if SSM, earlier and now it is no longer the GDR, then it sets the assert metric preference to maximum (0x7FFFFFFF) and the assert metric to one less than maximum (0xFFFFFFFE), as explained in [Section 5.7].

   If it was not the GDR for a group, or source and group pair if SSM, earlier, and the new hash does not make it GDR, then no processing is required.

   If it was not the GDR for an earlier group, or source and group pair if SSM, and now becomes the GDR, it starts building multicast forwarding tree for this flow.

3. If the router was included in the previous GDR list, but is not included in the new GDR list, or there is no new GDR list: The router MUST for each of the groups, or source and group pairs if the group is in SSM mode, with local receiver interest do as follows.

   If it was the GDR for a group, or source and group pair if SSM, it sets the assert metric preference to maximum (0x7FFFFFFF) and the assert metric to one less than maximum (0xFFFFFFFE), as explained in [Section 5.7].

   If it was not the GDR, then no processing is required.

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
(0xFFFFFFFFF). This was also mentioned in the previous section. 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 routing. This is similar to what is done for
AssertCancel Messages in [RFC7761], except that the metric value here
is one less.

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.

Using the above example, for G1, assume R1 and R3 agree on the new
GDR, which is R3. With the new 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.

For G2, assume it takes a slightly longer time for R2 to find out
that R3 is the new GDR and still considers itself being the GDR while
R3 already has assumed the role of GDR. Since both R2 and R3 think
they are GDRs, they further compare their metric and IP addresses.
If R3 has the better routing metric, or the same metric but a better
tie-breaker, the result will be consistent during GDR selection. If
unfortunately, R2 has the better metric or the same metric but a
better tie-breaker, R2 will become the Assert winner and continues to
forward traffic. Shortly after when R2 finds out that it is no
longer the GDR, R2 will change to using the near maximum assert
metric. Next time R2 sends an assert message, it will lose the

assert and stop forwarding. As assert winner, R2 would send periodic assert messages per [RFC7761].

5.8. Backward Compatibility

In the case of a hybrid Ethernet shared LAN (where some PIM routers enable the specification defined in this document, and some do not).

- If a router which does not support this specification becomes the DR on the LAN, then it is the only router acting as a DR, and there will be no load-balancing.

- If a router which does not support this specification becomes a non-DR on link, then it acts as non-DR defined in [RFC7761], and it will not take part in any load-balancing. Load-balancing may still happen.

6. Manageability Considerations

An administrator needs to consider what the total bandwidth requirements are and find a set of routers that together has enough total capacity, while making sure that each of the router 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. Ideally all the 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.

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Modulo</td>
<td>This document</td>
</tr>
<tr>
<td>1-255</td>
<td>Unassigned</td>
<td></td>
</tr>
</tbody>
</table>

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 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.

9. Acknowledgement

The authors would like to thank Steve Simlo and Taki Millonis for helping with the original idea; Alia Atlas, Bill Atwood, Jake Holland, Bharat Joshi, Anish Kachinthaya, Anvitha Kachinthaya and Alvaro Retana for reviews and comments; and Toerless Eckert and Rishabh Parekh for helpful conversation on the document.

10. References

10.1. Normative References


10.2. Informative References


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