Network Working Group
Request for Comments: 5059
Obsoletes: 2362
Updates: 4601
Category: Standards Track

N. Bhaskar
Arastra
A. Gall
SWITCH
J. Lingard
Arastra
S. Venaas
UNINETT

January 2008

Bootstrap Router (BSR) Mechanism for Protocol Independent Multicast (PIM)

Status of This Memo

Abstract

This document specifies the Bootstrap Router (BSR) mechanism for the class of multicast routing protocols in the PIM (Protocol Independent Multicast) family that use the concept of a Rendezvous Point as a means for receivers to discover the sources that send to a particular multicast group. BSR is one way that a multicast router can learn the set of group-to-RP mappings required in order to function. The mechanism is dynamic, largely self-configuring, and robust to router failure.

Table of Contents

1. Introduction

This document assumes some familiarity with the concepts of Protocol Independent Multicast - Sparse Mode (PIM-SM) [1] and Bidirectional Protocol Independent Multicast (BIDIR-PIM) [2], as well as with Administratively Scoped IP Multicast [3] and the IPv6 Scoped Address Architecture [4].

For correct operation, every multicast router within a PIM domain must be able to map a particular multicast group address to the same Rendezvous Point (RP). The PIM specifications do not mandate the use of a single mechanism to provide routers with the information to perform this group-to-RP mapping.

This document describes the PIM Bootstrap Router (BSR) mechanism. BSR is one way that a multicast router can learn the information required to perform the group-to-RP mapping. The mechanism is dynamic, largely self-configuring, and robust to router failure.

BSR was first defined in RFC 2362 [7] as part of the original PIM-SM specification, which has been obsoleted by RFC 4601 [1]. This document provides an updated specification of the BSR mechanism from RFC 2362, and also extends it to cope with administratively scoped region boundaries and different flavors of routing protocols.

Throughout the document, any reference to the PIM protocol family is restricted to the subset of RP-based protocols, namely PIM-SM and BIDIR-PIM, unless stated otherwise.

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 RFC 2119 [6].

1.1. Background

A PIM domain is a contiguous set of routers that all implement PIM and are configured to operate within a common boundary defined by PIM Multicast Border Routers (PMBRs). PMBRs connect each PIM domain to the rest of the Internet.

Every PIM multicast group needs to be associated with the IP address of a Rendezvous Point (RP). This address is used as the root of a group-specific distribution tree whose branches extend to all nodes in the domain that want to receive traffic sent to the group. Senders inject packets into the tree in such a manner that they reach all connected receivers. How this is done and how the packets are forwarded along the distribution tree depends on the particular routing protocol.

For all senders to reach all receivers, it is crucial that all routers in the domain use the same mappings of group addresses to RP addresses.

An exception to the above is where a PIM domain has been broken up into multiple administrative scope regions. These are regions where a border has been configured so that a set of multicast groups will not be forwarded across that border. In this case, all PIM routers within the same scope region must map a particular scoped group to the same RP within that region.

In order to determine the RP for a multicast group, a PIM router maintains a collection of group-to-RP mappings, called the RP-Set. A group-to-RP mapping contains the following elements.

- o Multicast group range, expressed as an address and prefix length
- o RP priority
- o RP address
- o Hash mask length
- o SM / BIDIR flag

In general, the group ranges of these group-to-RP mappings may overlap in arbitrary ways; hence, a particular multicast group may be covered by multiple group-to-RP mappings. When this is the case, the router chooses only one of the RPs by applying a deterministic algorithm so that all routers in the domain make the same choice. It is important to note that this algorithm is part of the specification of the individual routing protocols (and may differ among them), not of the BSR specification. For example, PIM-SM [1] defines one such algorithm. It makes use of a hash function for the case where a group range has multiple RPs with the same priority. The hash mask length is used by this function.

There are a number of ways in which such group-to-RP mappings can be established. The simplest solution is for all the routers in the domain to be statically configured with the same information. However, static configuration generally doesn't scale well, and, except when used in conjunction with Anycast-RP (see [8] and [9]), does not dynamically adapt to route around router or link failures.

The BSR mechanism provides a way in which viable group-to-RP mappings can be created and rapidly distributed to all the PIM routers in a domain. It is adaptive, in that if an RP becomes unreachable, this will be detected and the RP-Sets will be modified so that the unreachable RP is no longer used.

1.2. Protocol Overview

In this section we give an informal and non-definitive overview of the BSR mechanism. The definitive specification begins in section 2.

The general idea behind the BSR mechanism is that some of the PIM routers within a PIM domain are configured to be potential RPs for the domain. These are known as Candidate-RPs (C-RPs). A subset of the C-RPs will eventually be used as the actual RPs for the domain. In addition, some of the PIM routers in the domain are configured to be candidate bootstrap routers, or Candidate-BSRs (C-BSRs). One of these C-BSRs will be elected to be the bootstrap router (BSR) for the domain, and all the PIM routers in the domain will learn the result of this election through Bootstrap messages. The C-RPs will then report their candidacy to the elected BSR, which chooses a subset of these C-RPs and distributes corresponding group-to-RP mappings to all the routers in the domain through Bootstrap messages.

In more detail, the BSR mechanism works as follows. There are four basic phases (although in practice, all phases may be occurring simultaneously):

1. BSR Election. Each Candidate-BSR originates Bootstrap messages (BSMs). Every BSM contains a BSR Priority field. Routers within the domain flood the BSMs throughout the domain. A C-BSR that hears about a higher-priority C-BSR than itself suppresses its sending of further BSMs for some period of time. The single remaining C-BSR becomes the elected BSR, and its BSMs inform all the other routers in the domain that it is the elected BSR.

2. C-RP Advertisement. Each Candidate-RP within a domain sends periodic Candidate-RP-Advertisement (C-RP-Adv) messages to the elected BSR. A C-RP-Adv message includes the priority of the advertising C-RP, as well as a list of group ranges for which the candidacy is advertised. In this way, the BSR learns about possible RPs that are currently up and reachable.

3. RP-Set Formation. The BSR selects a subset of the C-RPs that it has received C-RP-Adv messages from to form the RP-Set. In general, it should do this in such a way that the RP-Set is neither so large that all the routers in the domain cannot be informed about it, nor so small that the load is overly concentrated on some RPs. It should also attempt to produce an RP-Set that does not change frequently.

4. RP-Set Flooding. In future Bootstrap messages, the BSR includes the RP-Set information. Bootstrap messages are flooded through the domain, which ensures that the RP-Set rapidly reaches all the routers in the domain. BSMs are originated periodically to ensure consistency after failure restoration.

   When a PIM router receives a Bootstrap message, it adds the group-to-RP mappings contained therein to its pool of mappings obtained from other sources (e.g., static configuration). It calculates the final mappings of group addresses to RP addresses from this pool according to rules specific to the particular routing protocol and uses that information to construct multicast distribution trees.

If a PIM domain becomes partitioned, each area separated from the old BSR will elect its own BSR, which will distribute an RP-Set containing RPs that are reachable within that partition. When the partition heals, another election will occur automatically and only one of the BSRs will continue to send out Bootstrap messages. As is expected at the time of a partition or healing, some disruption in packet delivery may occur. The duration of the disruption period will be on the order of the region's round-trip time and the BS_Timeout value.

1.3. Administrative Scoping and BSR

The mechanism described in the previous section does not work when the PIM domain is divided into administratively scoped regions. To handle this situation, we use the protocol modifications described in this section.

In the remainder of this document, we will use the term scope zone, or simply zone, when we are talking about a connected region of topology of a given scope. For a more precise definition of scope zones, see [4], which emphasizes that the scope zones are administratively configured.

Administrative scoping permits a PIM domain to be divided into multiple admin-scope zones. Each admin-scope zone is a convex connected set of PIM routers and is associated with a set of group addresses. The boundary of the admin-scope zone is formed by Zone Border Routers (ZBRs). ZBRs are configured not to forward traffic for any of the scoped group addresses into or out of the scoped zone. It is important to note that a given scope boundary always creates at least two scoped zones: one on either side of the boundary.

In IPv4, administratively scoped zones are associated with a set of addresses given by an address and a prefix length. In IPv6, administratively scoped zones are associated with a set of addresses given by a single scope ID value. The set of addresses corresponding to a given scope ID value is defined in [5]. For example, a scope ID of 5 maps to the 16 IPv6 address ranges ff[0-f]5::/16.

There are certain topological restrictions on admin-scope zones. The scope zone border must be complete and convex. By this we mean that there must be no path from the inside to the outside of the scoped zone that does not pass through a configured scope border router, and that the multicast capable path between any arbitrary pair of multicast routers in the scope zone must remain in the zone.

Administrative scoping complicates BSR because we do not want a PIM router within the scoped zone to use an RP outside the scoped zone. Thus we need to modify the basic mechanism to ensure that this doesn't happen.

This is done by running a separate copy of the basic BSR mechanism, as described in the previous section, within each admin-scope zone of a PIM domain. Thus a separate BSR election takes place for each admin-scope zone, a C-RP typically registers to the BSR of every admin-scope zone it is in, and every PIM router receives Bootstrap messages for every scope zone it is in. The Bootstrap messages sent by the BSR for a particular scope zone contain information about the RPs that should be used for the set of addresses associated with that scope zone.

Bootstrap messages are marked to indicate which scope zone they belong to. Such admin-scoped Bootstrap messages are flooded in the normal way, but will not be forwarded by a ZBR across the boundary for that scope zone.

For the BSR mechanism to function correctly with admin scoping, there must be at least one C-BSR within each admin-scope zone, and there must be at least one C-RP that is configured to be a C-RP for the set of group addresses associated with the scoped zone.

Even when administrative scoping is used, a copy of the BSR mechanism is still used across the entire PIM domain in order to distribute RP information for groups that are not administratively scoped. We call this copy of the mechanism non-scoped BSR. The copies of the mechanism run for each admin-scope zone are called scoped BSR.

Only the C-BSRs and the ZBRs need to be configured to know about the existence of the scope zones. Other routers, including the C-RPs, learn of their existence from Bootstrap messages.

All PIM routers within a PIM bootstrap domain where admin-scope ranges are in use must be capable of receiving Bootstrap messages and storing the winning BSR and RP-Set for all admin-scope zones that apply. Thus, PIM routers that only implement RFC 2362 or non-scoped BSR (which only allows one BSR per domain) cannot be used within the admin-scope zones of a PIM domain.

2. BSR State and Timers

A PIM router implementing BSR holds the following state.

RP-Set

Per Configured or Learned Scope Zone (Z):

- At all routers:
  - Current Bootstrap Router's IP Address
  - Current Bootstrap Router's BSR Priority
  - Last BSM received from current BSR
  - Bootstrap Timer (BST(Z))
  - Per group-to-RP mapping (M):
    - Group-to-RP mapping Expiry Timer (GET(M,Z))
- At a Candidate-BSR for Z:
  - My state: One of "Candidate-BSR", "Pending-BSR", "Elected-BSR"
- At a router that is not a Candidate-BSR for Z:
  - My state: One of "Accept Any", "Accept Preferred"
  - Scope-Zone Expiry Timer (SZT(Z))
- At the current Bootstrap Router for Z only:
  - Per group-to-C-RP mapping (M):
    - Group-to-C-RP mapping Expiry Timer (CGET(M,Z))

At a C-RP only:

C-RP Advertisement Timer (CRPT)

3. Bootstrap Router Election and RP-Set Distribution

3.1. Bootstrap Router Election

For simplicity, Bootstrap messages are used in both the BSR election and the RP-Set distribution mechanisms.

Each Bootstrap message indicates the scope to which it belongs. If the Admin Scope Zone bit is set in the first group range in the Bootstrap message, the message is called a scoped BSM. If the Admin Scope Zone bit is not set in the first group range in the Bootstrap message, the message is called a non-scoped BSM.

In a scoped IPv4 BSM, the scope of the message is given by the first group range in the message, which can be any sub-range of 224/4. In a scoped IPv6 BSM, the scope of the message is given by the scope ID of the first group range in the message, which must have a mask length of at least 16. For example, a group range of ff05::/16 with the Admin Scope Zone bit set indicates that the Bootstrap message is for the scope with scope ID 5. If the mask length of the first group range in a scoped IPv6 BSM is less than 16, the message MUST be dropped and a warning SHOULD be logged.

The state machine for Bootstrap messages depends on whether or not a router has been configured to be a Candidate-BSR for a particular scope zone. The per-scope-zone state machine for a C-BSR is given below, followed by the state machine for a router that is not configured to be a C-BSR.

A key part of the election mechanism is that we associate a weight with each BSR. The weight of a BSR is defined to be the concatenation in fixed-precision unsigned arithmetic of the BSR Priority field from the Bootstrap message and the IP address of the BSR from the Bootstrap message (with the BSR Priority taking the most-significant bits and the IP address taking the least-significant bits).

3.1.1. Per-Scope-Zone Candidate-BSR State Machine

When in C-BSR state			
Event	Receive Preferred BSM	Bootstrap Timer Expires	Receive Non-preferred BSM from Elected BSR
Action	-> C-BSR state Forward BSM; Store RP-Set; Set Bootstrap Timer to BS_Timeout	-> P-BSR state Set Bootstrap Timer to BS_Rand_Override	-> P-BSR state Forward BSM; Set Bootstrap Timer to BS_Rand_Override

When in P-BSR state			
Event	Receive Preferred BSM	Bootstrap Timer Expires	Receive Non-preferred BSM
Action	-> C-BSR state Forward BSM; Store RP-Set; Set Bootstrap Timer to BS_Timeout	-> E-BSR state Originate BSM; Set Bootstrap Timer to BS_Period	-> P-BSR state Forward BSM

When in E-BSR state			
Event	Receive Preferred BSM	Bootstrap Timer Expires	Receive Non-preferred BSM
Action	-> C-BSR state Forward BSM; Store RP-Set; Set Bootstrap Timer to BS_Timeout	-> E-BSR state Originate BSM; Set Bootstrap Timer to BS_Period	-> E-BSR state Originate BSM; Set Bootstrap Timer to BS_Period

A Candidate-BSR may be in one of three states for a particular scope zone:

Candidate-BSR (C-BSR)
The router is a candidate to be the BSR for the scope zone, but currently another router is the preferred BSR.

Pending-BSR (P-BSR)
The router is a candidate to be the BSR for the scope zone. Currently, no other router is the preferred BSR, but this router is not yet the elected BSR. This is a temporary state that prevents rapid thrashing of the choice of BSR during BSR election.

Elected-BSR (E-BSR)
The router is the elected BSR for the scope zone and it must perform all the BSR functions.

In addition to the three states, there is one timer:

o The Bootstrap Timer (BST) - used to time out old bootstrap router information, and used in the election process to terminate P-BSR state.

The initial state for this configured scope zone is "Pending-BSR"; the Bootstrap Timer is initialized to BS_Rand_Override. This is the case both if the router is a Candidate-BSR at startup, and if it is reconfigured to become one later.

3.1.2. Per-Scope-Zone State Machine for Non-Candidate-BSR Routers

The following state machine is used for scope zones that are discovered by the router from bootstrap messages. A simplified state machine is used for scope zones that are explicitly configured on the router and for the global zone. The differences are listed at the end of this section.

When in NoInfo state	
Event	Receive BSM
Action	-> AP state Forward BSM; Store RP-Set; Set Bootstrap Timer to BS_Timeout

When in Accept Any state		
Event	Receive BSM	Scope-Zone Expiry Timer Expires
Action	-> AP state Forward BSM; Store RP-Set; Set Bootstrap Timer to BS_Timeout	-> NoInfo state Remove scope zone state

When in Accept Preferred state			
Event	Receive Preferred BSM	Bootstrap Timer Expires	Receive Non-preferred BSM
Action	-> AP state Forward BSM; Store RP-Set; Set Bootstrap Timer to BS_Timeout	-> AA state Refresh RP-Set; Remove BSR state; Set SZT to SZ_Timeout	-> AP state

A router that is not a Candidate-BSR may be in one of three states:

NoInfo
The router has no information about this scope zone. When in this state, no state information is held and no timers (that refer to this scope zone) run. Conceptually, the state machine is only instantiated when the router receives a scoped BSM for a scope about which it has no prior knowledge. However, because the router immediately transitions to the AA state unconditionally, the NoInfo state can be considered to be virtual in a certain sense. For this reason, it is omitted from the description in section 2.

Accept Any (AA)
The router does not know of an active BSR, and will accept the first Bootstrap message it sees as giving the new BSR's identity and the RP-Set.

Accept Preferred (AP)
The router knows the identity of the current BSR, and is using the RP-Set provided by that BSR. Only Bootstrap messages from that BSR or from a C-BSR with higher weight than the current BSR will be accepted.

In addition to the three states, there are two timers:

- o The Bootstrap Timer (BST) - used to time out old bootstrap router information.

- o The Scope-Zone Expiry Timer (SZT) - used to time out the scope zone itself if Bootstrap messages specifying this scope zone stop arriving.

The initial state for scope zones about which the router has no knowledge is "NoInfo".

The state machine used for scopes that have been configured explicitly on the router and for the global scope (which always exists) differs from the state machine above as follows.

- o The "NoInfo" state doesn't exist.

- o No SZT is maintained. Hence, the event "Scope-Zone Expiry Timer Expires" does not exist and no actions with regard to this timer are executed.

The initial state for this state machine is "Accept Any".

3.1.3. Bootstrap Message Processing Checks

When a Bootstrap message is received, the following initial checks must be performed:

```
if ((DirectlyConnected(BSM.src_ip_address) == FALSE) OR
     (we have no Hello state for BSM.src_ip_address)) {
  drop the Bootstrap message silently
}

if (BSM.dst_ip_address == ALL-PIM-ROUTERS) {
  if (BSM.no_forward_bit == 0) {
    if (BSM.src_ip_address != RPF_neighbor(BSM.BSR_ip_address)) {
      drop the Bootstrap message silently
    }
  } else if ((any previous BSM for this scope has been accepted) OR
             (more than BS_Period has elapsed since startup)) {
```

```
        #only accept no-forward BSM if quick refresh on startup
        drop the Bootstrap message silently
      }
    } else if ((Unicast BSM support enabled) AND
               (BSM.dst_ip_address is one of my addresses)) {
      if ((any previous BSM for this scope has been accepted) OR
          (more than BS_Period has elapsed since startup)) {
        #the packet was unicast, but this wasn't
        #a quick refresh on startup
        drop the Bootstrap message silently
      }
    } else {
      drop the Bootstrap message silently
    }

    if (the interface the message arrived on is an admin scope
        border for the BSM.first_group_address) {
      drop the Bootstrap message silently
    }
```

Basically, the packet must have come from a directly connected neighbor for which we have active Hello state. It must have been sent to the ALL-PIM-ROUTERS group, and unless it is a No-Forward BSM, it must have been sent by the correct upstream router towards the BSR that originated the Bootstrap message; or, if it is a No-Forward BSM, we must have recently restarted and have no BSR state for that admin scope. Also, if unicast BSM support is enabled, a unicast BSM is accepted if it is addressed to us, we have recently restarted, and we have no BSR state for that admin scope. In addition, it must not have arrived on an interface that is a configured admin-scope border for the first group address contained in the Bootstrap message.

3.1.4. State Machine Transition Events

If the Bootstrap message passes the initial checks above without being discarded, then it may cause a state transition event in one of the above state machines. For both candidate and non-candidate BSRs, the following transition events are defined:

Receive Preferred BSM
: A Bootstrap message is received from a BSR that has weight higher than or equal to that of the current BSR. If a router is in P-BSR state, then it uses its own weight as that of the current BSR.

  A Bootstrap message is also preferred if it is from the current BSR with a lower weight than the previous BSM it sent, provided that if the router is a Candidate-BSR the current BSR still has a weight higher than or equal to that of the router itself. In this case, the "Current Bootstrap Router's BSR Priority" state must be updated. (For lower weight, see Non-preferred BSM from Elected BSR case.)

Receive Non-preferred BSM
: A Bootstrap message is received from a BSR other than the current BSR that has lower weight than that of the current BSR. If a router is in P-BSR state, then it uses its own weight as that of the current BSR.

Receive Non-preferred BSM from Elected BSR
: A Bootstrap message is received from the elected BSR, but the BSR Priority field in the received message has changed, so that now the currently elected BSR has lower weight than that of the router itself.

Receive BSM
: A Bootstrap message is received, regardless of BSR weight.

In addition to state machine transitions caused by the receipt of Bootstrap messages, a state machine transition takes place each time the Bootstrap Timer or Scope-Zone Expiry Timer expires.

3.1.5. State Machine Actions

The state machines specify actions that include setting the Bootstrap Timer and the Scope-Zone Expiry Timer to various values. These values are defined in section 5.

In addition to setting and cancelling the timers, the following actions may be triggered by state changes in the state machines:

Forward BSM
: A multicast Bootstrap message with No-Forward bit cleared that passes the Bootstrap Message Processing Checks is forwarded out of all interfaces with PIM neighbors (including the interface it is received on), except where this would cause the BSM to cross an admin-scope boundary for the scope zone indicated in the message. For details, see section 3.4.

Originate BSM
: A new Bootstrap message is constructed by the BSR, giving the BSR's address and BSR priority, and containing the BSR's chosen RP-Set. The message is forwarded out of all interfaces on which PIM neighbors exist, except where this would cause the BSM to cross an admin-scope boundary for the scope zone indicated in the message.

Store RP-Set

The router uses the group-to-RP mappings contained in a BSM to update its local RP-Set.

This action is skipped for an empty BSM. A BSM is empty if it contains no group ranges, or if it only contains a single group range where that group range has the Admin Scope Zone bit set (a scoped BSM) and an RP count of zero.

If a mapping does not yet exist, it is created and the associated Group-to-RP mapping Expiry Timer (GET) is initialized with the holdtime from the BSM.

If a mapping already exists, its GET is set to the holdtime from the BSM. If the holdtime is zero, the mapping is removed immediately. Note that for an existing mapping, the RP priority must be updated if changed.

Mappings for a group range are also to be immediately removed if they are not present in the received group range. This means that if there are any existing group-to-RP mappings for a range where the respective RPs are not in the received range, then those mappings must be removed.

All RP mappings associated with the scope zone of the BSM are updated with the new hash mask length from the received BSM. This includes RP mappings for all group ranges learned for this zone, not just the ranges in this particular BSM.

In addition, the entire BSM is stored for use in the action Refresh RP-Set and to prime a new PIM neighbor as described below.

Refresh RP-Set

When the Bootstrap Timer expires, the router uses the copy of the last BSM that it has received to refresh its RP-Set according to the action Store RP-Set as if it had just received it. This will increase the chance that the group-to-RP mappings will not expire during the election of the new BSR.

Remove BSR state

When the Bootstrap Timer expires, all state associated with the current BSR is removed (address, priority, BST, and saved last BSM; see section 2). Note that this does not include any group-to-RP mappings.

Remove scope zone state
When the Scope-Zone Expiry Timer expires, all state associated with the scope zone is removed (see section 2).

3.2. Sending Candidate-RP-Advertisement Messages

Every C-RP periodically unicasts a C-RP-Adv message to the BSR for each scope zone for which it has state, to inform the BSR of the C-RP's willingness to function as an RP. These messages are sent with an interval of C_RP_Adv_Period, except when a new BSR is elected; see below.

When a new BSR is elected, the C-RP MUST send one to three C-RP-Adv messages and wait a small randomized period C_RP_Adv_Backoff before sending each message. We recommend sending three messages because it is important that the BSR quickly learns which RPs are active, and some packet loss may occur when a new BSR is elected due to changes in the network. One way of implementing this is to set the CRPT to C_RP_Adv_Backoff when the new BSR is elected, as well as setting a counter to 2. Whenever the CRPT expires, we first send a C-RP-Adv message as usual. Next, if the counter is non-zero, it is decremented and the CRPT is again set to C_RP_Adv_Backoff instead of C_RP_Adv_Period.

The Priority field in these messages is used by the BSR to select which C-RPs to include in the RP-Set. Note that lower values of this field indicate higher priorities, so that a value of zero is the highest possible priority. C-RPs should, by default, send C-RP-Adv messages with the Priority field set to 192.

When a C-RP is being shut down, it SHOULD immediately send a C-RP-Adv message to the BSR for each scope zone for which it is currently serving as an RP; the Holdtime in this C-RP-Adv message should be zero. The BSR will then immediately time out the C-RP and generate a new Bootstrap message with the shut down RP holdtime set to 0.

A C-RP-Adv message carries a list of group address and group mask field pairs. This enables the C-RP to specify the group ranges for which it is willing to be the RP. If the C-RP becomes an RP, it may enforce this scope acceptance when receiving Register or Join/Prune messages.

A C-RP is configured with a list of group ranges for which it should advertise itself as the C-RP. A C-RP uses the following algorithm to determine which ranges to send to a given BSR.

For each group range R in the list, the C-RP advertises that range to the scoped BSR for the smallest scope that "contains" R. For IPv6, the containing scope is determined by matching the scope identifier of the group range with the scope of the BSR. For IPv4, it is the longest-prefix match for R, amongst the known admin-scope ranges. If no scope is found to contain the group range, the C-RP includes it in the C-RP-Adv sent to the non-scoped BSR. If a non-scoped BSR is not known, the range is not included in any C-RP-Adv.

In addition, for each IPv4 group range R in the list, for each scoped BSR whose scope range is strictly contained within R, the C-RP SHOULD by default advertise that BSR's scope range to that BSR. And for each IPv6 group range R in the list with prefix length < 16, the C-RP SHOULD by default advertise each sub-range of prefix length 16 to the scoped BSR with the corresponding scope ID. An implementation MAY supply a configuration option to prevent the behavior described in this paragraph, but such an option SHOULD be disabled by default.

For IPv6, the mask length of all group ranges included in the C-RP-Adv message sent to a scoped BSR MUST be >= 16.

If the above algorithm determines that there are no group ranges to advertise to the BSR for a particular scope zone, a C-RP-Adv message MUST NOT be sent to that BSR. A C-RP MUST NOT send a C-RP-Adv message with no group ranges in it.

If the same router is the BSR for more than one scope zone, the C-RP-Adv messages for these scope zones MAY be combined into a single message.

If the C-RP is a ZBR for an admin-scope zone, then the Admin Scope Zone bit MUST be set in the C-RP-Adv messages it sends for that scope zone; otherwise this bit MUST NOT be set. This information is currently only used for logging purposes by the BSR, but might allow for future extensions of the protocol.

3.3. Creating the RP-Set at the BSR

Upon receiving a C-RP-Adv message, the router needs to decide whether or not to accept each of the group ranges included in the message. For each group range in the message, the router checks to see if it is the elected BSR for any scope zone that contains the group range, or if it is elected as the non-scoped BSR. If so, the group range is accepted; if not, the group range is ignored.

For security reasons, we recommend that implementations have a way of restricting which IP addresses the BSR accepts C-RP-Adv messages from, e.g., access lists. For use of scoped BSR, it may also be useful to specify which group ranges should be accepted.

If the group range is accepted, a group-to-C-RP mapping is created for this group range and the RP Address from the C-RP-Adv message.

If the mapping is not already part of the C-RP-Set, it is added to the C-RP-Set and the associated Group-to-C-RP mapping Expiry Timer (CGET) is initialized to the holdtime from the C-RP-Adv message. Its priority is set to the Priority from the C-RP-Adv message.

If the mapping is already part of the C-RP-Set, it is updated with the Priority from the C-RP-Adv message, and its associated CGET is reset to the holdtime from the C-RP-Adv message. If the holdtime is zero, the mapping is immediately removed from the C-RP-Set.

The hash mask length is a global property of the BSR and is therefore the same for all mappings managed by the BSR.

For compatibility with the previous version of the BSR specification, a C-RP-Adv message with no group ranges SHOULD be treated as though it contained the single group range ff00::/8 or 224/4. Therefore, according to the rule above, this group range will be accepted if and only if the router is elected as the non-scoped BSR.

When a CGET expires, the corresponding group-to-C-RP mapping is removed from the C-RP-Set.

The BSR constructs the RP-Set from the C-RP-Set. It may apply a local policy to limit the number of Candidate-RPs included in the RP-Set. The BSR may override the range indicated in a C-RP-Adv message unless the 'Priority' field from the C-RP-Adv message is less than 128.

If the BSR learns of both BIDIR and PIM-SM Candidate-RPs for the same group range, the BSR MUST only include RPs for one of the protocols in the BSMs. The default behavior SHOULD be to prefer BIDIR.

For inclusion in a BSM, the RP-Set is subdivided into sets of {group-range, RP-Count, RP-addresses}. For each RP-address, the "RP-Holdtime" field is set to the Holdtime from the C-RP-Set, subject to the constraint that it MUST be larger than BS_Period and SHOULD be larger than 2.5 times BS_Period to allow for some Bootstrap messages getting lost. If some holdtimes from the C-RP-Sets do not satisfy this constraint, the BSR MUST replace those holdtimes with a value satisfying the constraint. An exception to this is the holdtime of zero, which is used to immediately withdraw mappings.

The format of the Bootstrap message allows 'semantic fragmentation', if the length of the original Bootstrap message exceeds the packet maximum boundaries. However, to reduce the semantic fragmentation required, we recommend against configuring a large number of routers as C-RPs.

In general, BSMs are originated at regular intervals according to the BS_Period timer. We do recommend that a BSM is also originated whenever the RP-set to be announced in the BSMs changes. This will usually happen when receiving C-RP advertisements from a new C-RP, or when a C-RP is shut down (C-RP advertisement with a holdtime of zero). There MUST however be a minimum of BS_Min_Interval between each time a BSM is sent. In particular, when a new BSR is elected, it will first send one BSM (which is likely to be empty since it has not yet received any C-RP advertisements), and then wait at least BS_Min_Interval before sending a new one. During that time, it is likely to have received C-RP advertisements from all usable C-RPs (since we say that a C-RP should send one or more advertisements with small random delays of C_RP_Adv_Backoff when a new BSR is elected). For this case in particular, where routers may not have a usable RP-set, we recommend originating a BSM as soon as BS_Min_Interval has passed. We suggest though that a BSR can do this in general. One way of implementing this, is to decrease the Bootstrap Timer to BS_Min_Interval whenever the RP-set changes, while not changing the timer if it is less than or equal to BS_Min_Interval.

A BSR originates separate scoped BSMs for each scope zone for which it is the elected BSR, as well as originating non-scoped BSMs if it is the elected non-scoped BSR.

Each group-to-C-RP mapping is included in precisely one of these BSMs -- namely, the scoped BSM for the narrowest scope containing the group range of the mapping, if any, or the non-scoped BSM otherwise.

A scoped BSM MUST have at least one group range, and the first group range in a scoped BSM MUST have the Admin Scope Zone bit set. This group range identifies the scope of the BSM. In a scoped IPv4 BSM, the first group range is the range corresponding to the scope of the BSM. In a scoped IPv6 BSM, the first group range may be any group range subject to the general condition that all the group ranges in such a BSM MUST have a mask length of at least 16 and MUST have the same scope ID as the scope of the BSM.

Apart from identifying the scope, the first group range in a scoped BSM is treated like any other range with respect to RP mappings. That is, all mappings in the RP-set for this group range, if any, must be included in this first group range in the BSM. After this group range, other group ranges in this scope (for which there are RP mappings) appear in any order.

The Admin Scope Zone bit of all group ranges other than the first SHOULD be set to 0 on origination, and MUST be ignored on receipt.

When an elected BSR is being shut down, it should immediately originate a Bootstrap message listing its current RP-Set, but with the BSR Priority field set to the lowest priority value possible. This will cause the election of a new BSR to happen more quickly.

3.4. Forwarding Bootstrap Messages

Generally, bootstrap messages originate at the BSR, and are hop-by-hop forwarded by intermediate routers if they pass the Bootstrap Message Processing Checks. There are two exceptions to this. One is that a bootstrap message is not forwarded if its No-Forward bit is set; see section 3.5.1. The other is that unicast BSMs (see section 3.5.2) are usually not forwarded. Implementers MAY, however, at their own discretion choose to re-send a No-Forward or unicast BSM in a multicast BSM, which MUST have the No-Forward bit cleared. It is essential that the No-Forward bit is cleared, since no Reverse Path Forwarding (RPF) check is performed by the receiver when it is set.

By hop-by-hop forwarding, we mean that the Bootstrap message itself is forwarded, not the entire IP packet. Each hop constructs an IP packet for each of the interfaces the BSM is to be forwarded out of; each packet contains the entire BSM that was received.

When a Bootstrap message is forwarded, it is forwarded out of every multicast-capable interface that has PIM neighbors (including the one over which the message was received). The exception to this is if the interface is an admin-scope boundary for the admin-scope zone indicated in the first group range in the Bootstrap message packet.

As an optimization, a router MAY choose not to forward a BSM out of the interface the message was received on if that interface is a point-to-point interface. On interfaces with multiple PIM neighbors, a router SHOULD forward an accepted BSM out of the interface that BSM was received on, but if the number of PIM neighbors on that interface is large, it MAY delay forwarding a BSM out of that interface by a small randomized interval to prevent message implosion. A configuration option MAY be provided to disable forwarding out of the interface a message was received on, but we recommend that the default behavior is to forward out of that interface.

Rationale: A BSM needs to be forwarded out of the interface the message was received on (in addition to the other interfaces) because the routers on a LAN may not have consistent routing information. If three routers on a LAN are A, B, and C, and at router B RPF(BSR)==A and at router C RPF(BSR)==B, then router A originally forwards the BSM onto the LAN, but router C will only accept it when router B re-forwards the message onto the LAN. If the underlying routing protocol configuration guarantees that the routers have consistent routing information, then forwarding out of the incoming interface may safely be disabled.

A ZBR constrains all BSMs that are of equal or smaller scope than the configured boundary. That is, the BSMs are not accepted from, originated, or forwarded on the interfaces on which the boundary is configured. For IPv6, the check is a comparison between the scope of the first range in the scoped BSM and the scope of the configured boundary. For IPv4, the first range in the scoped BSM is checked to see if it is contained in or is the same as the range of the configured boundary.

3.5. Bootstrap Messages to New and Rebooting Routers

When a Hello message is received from a new neighbor, or a Hello message with a new GenID is received from an existing neighbor, one router on the LAN sends a stored copy of the Bootstrap message for each admin-scope zone to the new or rebooting router. This allows new or rebooting routers to learn the RP-Set quickly.

This message SHOULD be sent as a No-Forward Bootstrap message; see section 3.5.1. For backwards compatibility, this message MAY instead or in addition be sent as a unicast Bootstrap message; see section 3.5.2. These messages MUST only be accepted at startup; see section 3.1.3.

The router that does this is the Designated Router (DR) on the LAN, or, if the new or rebooting router is the DR, the router that would be the DR if the new or rebooting router were excluded from the DR election process.

Before sending a Bootstrap message in this manner, the router must wait until it has sent a triggered Hello message on this interface; otherwise, the new neighbor will discard the Bootstrap message.

3.5.1. No-Forward Bootstrap Messages

A No-Forward Bootstrap message, is a bootstrap message that has the No-Forward bit set. All implementations SHOULD support sending of No-Forward Bootstrap messages, and SHOULD also accept them. The RPF check MUST NOT be performed in the BSM processing check for a No-Forward BSM; see section 3.1.3. The messages have the same source and destination addresses as the usual multicast Bootstrap messages.

3.5.2. Unicasting Bootstrap Messages

For backwards compatibility, implementations MAY support unicast Bootstrap messages. Whether to send unicast Bootstrap messages instead of or in addition to No-Forward Bootstrap messages, and also whether to accept such messages, SHOULD be configurable. This message is unicast to the neighbor.

3.6. Receiving and Using the RP-Set

The RP-Set maintained by BSR is used by RP-based multicast routing protocols like PIM-SM and BIDIR-PIM. These protocols may obtain RP-Sets from other sources as well. How the final group-to-RP mappings are obtained from these RP-Sets is not part of the BSR specification. In general, the routing protocols need to re-calculate the mappings when any of their RP-Sets change. How such a change is signalled to the routing protocol is also not part of the present specification.

Some group-to-RP mappings in the RP-Set indicate group ranges for which PIM-SM should be used; others indicate group ranges for use with BIDIR-PIM. Routers that support only one of these protocols MUST NOT ignore ranges indicated as being for the other protocol. They MUST NOT treat them as being for the protocol they support.

If a mapping is not already part of the RP-Set, it is added to the RP-Set and the associated Group-to-RP mapping Expiry Timer (GET) is initialized to the holdtime from the Bootstrap message. Its priority is set to the Priority from the Bootstrap message.

If a mapping is already part of the RP-Set, it is updated with the Priority from the Bootstrap message and its associated GET is reset to the holdtime from the Bootstrap message. If the holdtime is zero, the mapping is removed from the RP-Set immediately.

4. Message Formats

BSR messages are PIM messages, as defined in [1]. The values of the PIM Message Type field for BSR messages are:

4 Bootstrap

8 Candidate-RP-Advertisement

As with all other PIM control messages, BSR messages have IP protocol number 103.

Candidate-RP-Advertisement messages are unicast to a BSR. Usually, Bootstrap messages are multicast with TTL 1 to the ALL-PIM-ROUTERS group, but in some circumstances (described in section 3.5.2) Bootstrap messages may be unicast to a specific PIM neighbor.

The IP source address used for Candidate-RP-Advertisement messages is a domain-wide reachable address. The IP source address used for Bootstrap messages (regardless of whether they are being originated or forwarded) is the link-local address of the interface on which the message is being sent (i.e., the same source address that the router uses for the Hello messages that it sends out that interface).

The IPv4 ALL-PIM-ROUTERS group is 224.0.0.13. The IPv6 ALL-PIM-ROUTERS group is ff02::d.

In this section, we use the following terms defined in the PIM-SM specification [1]:

- o Encoded-Unicast format

- o Encoded-Group format

We repeat these here to aid readability.

Encoded-Unicast address

An Encoded-Unicast address takes the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Addr Family  | Encoding Type |     Unicast Address
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...
```

Addr Family
The PIM address family of the 'Unicast Address' field of this address.

Values of 0-127 are as assigned by the IANA for Internet Address Families in [11]. Values 128-250 are reserved to be assigned by the IANA for PIM-specific Address Families. Values 251 though 255 are designated for private use. As there is no assignment authority for this space, collisions should be expected.

Encoding Type
The type of encoding used within a specific Address Family. The value '0' is reserved for this field, and represents the native encoding of the Address Family.

Unicast Address
The unicast address as represented by the given Address Family and Encoding Type.

Encoded-Group address

Encoded-Group addresses take the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Addr Family  | Encoding Type |B| Reserved  |Z|  Mask Len     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Group multicast Address
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...
```

Addr Family
Described above.

Encoding Type
Described above.

[B]IDIR bit
When set, all BIDIR-capable PIM routers will operate the protocol described in [2] for the specified group range.

Reserved
Transmitted as zero. Ignored upon receipt.

Admin Scope [Z]one
When set, this bit indicates that this group range is an administratively scoped range.

Mask Len
The Mask length field is 8 bits. The value is the number of contiguous one bits that are left justified and used as a mask; when combined with the group address, it describes a range of groups. It is less than or equal to the address length in bits for the given Address Family and Encoding Type. If the message is sent for a single group, then the Mask length must equal the address length in bits for the given Address Family and Encoding Type (e.g., 32 for IPv4 native encoding and 128 for IPv6 native encoding).

Group multicast Address
Contains the group address.

4.1. Bootstrap Message Format

A Bootstrap message may be divided up into 'semantic fragments' if the resulting IP datagram would exceed the maximum packet size boundaries. Basically, a single Bootstrap message can be sent as multiple semantic fragments (each in a separate IP datagram), so long as the fragment tags of all the semantic fragments comprising the message are the same. The format of a single non-fragmented message is the same as the one used for semantic fragments.

The format of a single 'fragment' is given below:

```
     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
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |PIM Ver| Type  |N|  Reserved   |           Checksum            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |         Fragment Tag          | Hash Mask Len | BSR Priority  |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |             BSR Address (Encoded-Unicast format)              |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |            Group Address 1 (Encoded-Group format)             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | RP Count 1    | Frag RP Cnt 1 |         Reserved              |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |             RP Address 1 (Encoded-Unicast format)             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |          RP1 Holdtime         | RP1 Priority  |   Reserved    |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |             RP Address 2 (Encoded-Unicast format)             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |          RP2 Holdtime         | RP2 Priority  |   Reserved    |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                               .                               |
    |                               .                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |             RP Address m (Encoded-Unicast format)             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |          RPm Holdtime         | RPm Priority  |   Reserved    |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |            Group Address 2 (Encoded-Group format)             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                               .                               |
    |                               .                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |            Group Address n (Encoded-Group format)             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | RP Count n    | Frag RP Cnt n |          Reserved             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |             RP Address 1 (Encoded-Unicast format)             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |          RP1 Holdtime         | RP1 Priority  |   Reserved    |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |             RP Address 2 (Encoded-Unicast format)             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |          RP2 Holdtime         | RP2 Priority  |   Reserved    |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                               .                               |
    |                               .                               |
```

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             RP Address m (Encoded-Unicast format)             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          RPm Holdtime         | RPm Priority  |   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

PIM Version, Reserved, Checksum
Described in [1].

Type
PIM Message Type. Value is 4 for a Bootstrap message.

[N]o-Forward bit
When set, this bit means that the Bootstrap message fragment is not to be forwarded.

Fragment Tag
A randomly generated number, acts to distinguish the fragments belonging to different Bootstrap messages; fragments belonging to same Bootstrap message carry the same 'Fragment Tag'.

Hash Mask Len
The length (in bits) of the mask to use in the hash function. For IPv4, we recommend a value of 30. For IPv6, we recommend a value of 126.

BSR Priority
Contains the BSR priority value of the included BSR. This field is considered as a high-order byte when comparing BSR addresses. BSRs should by default set this field to 64. Note that for historical reasons, the highest BSR priority is 255 (the higher the better), whereas the highest RP Priority (see below) is 0 (the lower the better).

BSR Address
The address of the bootstrap router for the domain. The format for this address is given in the Encoded-Unicast address in [1].

Group Address 1..n
The group ranges (address and mask) with which the Candidate-RPs are associated. Format described in [1]. In a fragment containing admin-scope ranges, the first group range in the fragment MUST satisfy the following conditions:

- o it MUST have the Admin Scope Zone bit set;
- o for IPv4, it MUST be the group range for the entire admin-scope range (this is required even if there are no RPs in the RP-Set for the entire admin-scope range -- in this case, the sub-ranges for the RP-Set are specified later in the fragment along with their RPs);
- o for IPv6, the Mask Len MUST be at least 16 and have the scope ID of the admin-scope range.

RP Count 1..n
The number of Candidate-RP addresses included in the whole Bootstrap message for the corresponding group range. A router does not replace its old RP-Set for a given group range until/unless it receives 'RP-Count' addresses for that range; the addresses could be carried over several fragments. If only part of the RP-Set for a given group range was received, the router discards it without updating that specific group range's RP-Set.

Frag RP Cnt 1..m
The number of Candidate-RP addresses included in this fragment of the Bootstrap message, for the corresponding group range. The 'Frag RP Cnt' field facilitates parsing of the RP-Set for a given group range, when carried over more than one fragment.

RP address 1..m
The address of the Candidate-RPs, for the corresponding group range. The format for these addresses is given in the Encoded-Unicast address in [1].

RP1..m Holdtime
The Holdtime (in seconds) for the corresponding RP. This field is copied from the 'Holdtime' field of the associated RP stored at the BSR.

RP1..m Priority
The 'Priority' of the corresponding RP and Encoded-Group Address. This field is copied from the 'Priority' field stored at the BSR when receiving a C-RP-Adv message. The highest priority is '0' (i.e., unlike BSR priority, the lower the value of the 'Priority' field, the better). Note that the priority is per RP and per Group Address.

Within a Bootstrap message, the BSR Address, all the Group Addresses, and all the RP Addresses MUST be of the same address family. In addition, the address family of the fields in the message MUST be the same as the IP source and destination addresses of the packet. This permits maximum implementation flexibility for dual-stack IPv4/IPv6 routers.

4.1.1. Semantic Fragmentation of BSMs

Bootstrap messages may be split over several PIM Bootstrap Message Fragments (BSMFs); this is known as semantic fragmentation. Each of these must follow the above format. All fragments of a given Bootstrap message MUST have identical values for the Type, No-Forward bit, Fragment Tag, Hash Mask Len, BSR Priority, and BSR Address fields. That is, only the group-to-RP mappings may differ between fragments.

This is useful if the BSM would otherwise exceed the MTU of the link the message will be forwarded over. If one relies purely on IP fragmentation, one would lose the entire message if a single fragment is lost. By use of semantic fragmentation, a single lost IP fragment will only cause the loss of the semantic fragment that the IP fragment was part of. As described below, a router only needs to receive all the RPs for a specific group range to update that range. This means that loss of a semantic fragment, due to an IP fragment getting lost, only affects the group ranges for which the lost semantic fragment contains information.

If the BSR can split up the BSM so that each group range (and all of its RP information) can fit entirely inside one BSMF, then it should do so. If a BSMF is lost, the state from the previous BSM for the group ranges from the missing BSMF will be retained. Each fragment that does arrive will update the RP information for the group ranges contained in that fragment, and the new group-to-RP mappings for those can be used immediately. The information from the missing fragment will be obtained when the next BSM is transmitted.

If the list of RPs for a single group range is long, one may split the information across multiple BSMFs to avoid IP fragmentation. In this case, all the BSMFs comprising the information for that group range must be received before the group-to-RP mapping in use can be modified. This is the purpose of the RP Count field -- a router receiving BSMFs from the same BSM (i.e., that have the same fragment tag) must wait until BSMFs providing RP Count RPs for that group range have been received before the new group-to-RP mapping can be used for that group range. If a single BSMF from such a large group range is lost, then that entire group range will have to wait until the next BSM is originated. Hence, in this case, the benefit of using semantic fragmentation is dubious.

Next we need to consider how a BSR would remove group ranges. A router receiving a set of BSMFs cannot tell if a group range is missing. If it has seen a group range before, it must assume that that group range still exists, and that the BSMF describing that group range has been lost. The router should retain this information for BS_Timeout. Thus, for a BSR to remove a group range, it should include that group range, but with an RP Count of zero, and it should resend this information in each BSM for BS_Timeout.

4.2. Candidate-RP-Advertisement Message Format

Candidate-RP-Advertisement messages are periodically unicast from the C-RPs to the BSR.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver| Type  |   Reserved    |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Prefix Count  |   Priority    |           Holdtime            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             RP Address (Encoded-Unicast format)               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Group Address 1 (Encoded-Group format)             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               .                               |
|                               .                               |
|                               .                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Group Address n (Encoded-Group format)             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

PIM Version, Reserved, Checksum
    Described in [1].

Type
    PIM Message Type. Value is 8 for a Candidate-RP-Advertisement message.

Prefix Count
    The number of Encoded-Group Addresses included in the message; indicating the group range for which the C-RP is advertising. C-RPs MUST NOT send C-RP-Adv messages with a Prefix Count of '0'.

Priority
    The 'Priority' of the included RP, for the corresponding Encoded- Group Address (if any). The highest priority is '0' (i.e., the lower the value of the 'Priority' field, the higher the priority). This field is stored at the BSR upon receipt along with the RP address and corresponding Encoded-Group Address.

Holdtime
    The amount of time (in seconds) the advertisement is valid. This field allows advertisements to be aged out. This field should be set to 2.5 times C_RP_Adv_Period.

RP Address
    The address of the interface to advertise as a Candidate-RP. The format for this address is given in the Encoded-Unicast address in [1].

Group Address-1..n
    The group ranges for which the C-RP is advertising. Format described in Encoded-Group-Address in [1].

Within a Candidate-RP-Advertisement message, the RP Address and all the Group Addresses MUST be of the same address family. In addition, the address family of the fields in the message MUST be the same as the IP source and destination addresses of the packet. This permits maximum implementation flexibility for dual-stack IPv4/IPv6 routers.

5. Timers and Timer Values

Timer Name: Bootstrap Timer (BST(Z))

Value Name	Value	Explanation
BS_Period	Default: 60 seconds	Periodic interval with which BSMs are normally originated
BS_Timeout	Default: 130 seconds	Interval after which a BSR is timed out if no BSM is received from that BSR
BS_Min_Interval	Default: 10 seconds	Minimum interval with which BSMs may be originated
BS_Rand_Override	see below	Randomized interval used to reduce control message overhead during BSR election

Note that BS_Timeout MUST be larger than BS_Period, even if their values are changed from the defaults. We recommend that BS_Timeout is set to 2 times BS_Period plus 10 seconds.

BS_Rand_Override is calculated using the following pseudocode, in which all values are in units of seconds. The values of BS_Rand_Override generated by this pseudocode are between 5 and 23 seconds, with smaller values generated if the C-BSR has a high bootstrap weight, and larger values generated if the C-BSR has a low bootstrap weight.

```
BS_Rand_Override = 5 + priorityDelay + addrDelay
```

where priorityDelay is given by:

```
priorityDelay = 2 * log_2(1 + bestPriority - myPriority)
```

and addrDelay is given by the following for IPv4:

```
if (bestPriority == myPriority) {
    addrDelay = log_2(1 + bestAddr - myAddr) / 16
} else {
    addrDelay = 2 - (myAddr / 2^31)
}
```

and addrDelay is given by the following for IPv6:

```
if (bestPriority == myPriority) {
    addrDelay = log_2(1 + bestAddr - myAddr) / 64
} else {
    addrDelay = 2 - (myAddr / 2^127)
}
```

and bestPriority is given by:

```
bestPriority = max(storedPriority, myPriority)
```

and bestAddr is given by:

```
bestAddr = max(storedAddr, myAddr)
```

and where myAddr is the Candidate-BSR's address, storedAddr is the stored BSR's address, myPriority is the Candidate-BSR's configured priority, and storedPriority is the stored BSR's priority.

Timer Name: Scope Zone Expiry Timer (SZT(Z))

Value Name	Value	Explanation
SZ_Timeout	Default: 1300 seconds	Interval after which a scope zone is timed out if no BSM is received for that scope zone

Note that SZ_Timeout MUST be larger than BS_Timeout, even if their values are changed from the defaults. We recommend that SZ_Timeout is set to 10 times BS_Timeout.

Timer Name: Group-to-C-RP mapping Expiry Timer (CGET(M,Z))

Value Name	Value	Explanation
C-RP Mapping Timeout	from message	Holdtime from C-RP-Adv message

Timer Name: Group-to-RP mapping Expiry Timer (GET(M,Z))

Value Name	Value	Explanation
RP Mapping Timeout	from message	Holdtime from BSM

Timer Name: C-RP Advertisement Timer (CRPT)

Value Name	Value	Explanation
C_RP_Adv_Period	Default: 60 seconds	Periodic interval with which C-RP-Adv messages are sent to a BSR
C_RP_Adv_Backoff	Default: 0-3 seconds	Whenever a triggered C_RP_Adv is sent, a new randomized value between 0 and 3 is used

6. Security Considerations

6.1. Possible Threats

Threats affecting the PIM BSR mechanism are primarily of two forms: denial-of-service (DoS) attacks and traffic-diversion attacks. An attacker that subverts the BSR mechanism can prevent multicast traffic from reaching the intended recipients, can divert multicast traffic to a place where they can monitor it, and can potentially flood third parties with traffic.

Traffic can be prevented from reaching the intended recipients by one of two mechanisms:

- o Subverting a BSM, and specifying RPs that won't actually forward traffic.

- o Registering with the BSR as a C-RP, and then not forwarding traffic.

Traffic can be diverted to a place where it can be monitored by both of the above mechanisms; in this case, the RPs would forward the traffic, but are located so as to aid monitoring or man-in-the-middle attacks on the multicast traffic.

A third party can be flooded by either of the above two mechanisms by specifying the third party as the RP, and register traffic will then be forwarded to the third party.

6.2. Limiting Third-Party DoS Attacks

The third-party DoS attack above can be greatly reduced if PIM routers acting as DR do not continue to forward Register traffic to the RP in the presence of ICMP Protocol Unreachable or ICMP Host Unreachable responses. If a PIM router sending Register packets to an RP receives one of these responses to a data packet it has sent, it should rate- limit the transmission of future Register packets to that RP for a short period of time.

As this does not affect interoperability, the precise details are left to the implementer to decide. However, we note that a router implementing such rate limiting must only do so if the ICMP packet correctly echoes part of a Register packet that was sent to the RP. If this check were not made, then simply sending ICMP Unreachable packets to the DR with the source address of the RP spoofed would be sufficient to cause a denial-of-service attack on the multicast traffic originating from that DR.

6.3. Bootstrap Message Security

If a legitimate PIM router in a domain is compromised, there is little any security mechanism can do to prevent that router from subverting PIM traffic in that domain.

Implementations SHOULD provide a per-interface configuration option where one can specify that no Bootstrap messages are to be sent out of or accepted on the interface. This should generally be configured on all PMBRs in order not to receive messages from neighboring domains. This avoids receiving legitimate messages with conflicting BSR information from other domains, and also prevents BSR attacks from neighboring domains. This option is also useful on leaf interfaces where there are only hosts present. However, the Security Considerations section of [1] states that there should be a mechanism for not accepting PIM Hello messages on leaf interfaces and that messages should only be accepted from valid PIM neighbors. There may however be additional issues with unicast Bootstrap messages; see below. In addition to dropping all multicast Bootstrap messages on PMBRs, we also recommend configuring PMBRs (both towards other domains and on leaf interfaces) to drop all unicast PIM messages (Bootstrap message, Candidate-RP Advertisement, PIM register, and PIM register stop).

6.3.1. Unicast Bootstrap Messages

There are some possible security issues with unicast Bootstrap messages. The Bootstrap Message Processing Checks prevent a router from accepting a Bootstrap message from outside of the PIM Domain, as the source address on Bootstrap messages must be an immediate PIM neighbor. There is however a small window of time after a reboot where a PIM router will accept a bad Bootstrap message that is unicast from an immediate neighbor, and it might be possible to unicast a Bootstrap message to a router during this interval from outside the domain, using the spoofed source address of a neighbor. The best way to protect against this is to use the above-mentioned mechanism of configuring border and leaf interfaces to drop all bootstrap messages, including unicast messages. This can also be prevented if PMBRs perform source-address filtering to prevent packets entering the PIM domain with IP source addresses that are infrastructure addresses in the PIM domain.

The use of unicast Bootstrap messages is for backwards compatibility only. Due to the possible security implications, implementations supporting unicast Bootstrap messages SHOULD provide a configuration option for whether they are to be used.

6.3.2. Multi-Access Subnets

As mentioned above, implementations SHOULD provide a per-interface configuration option so that leaf interfaces and interfaces facing other domains can be configured to drop all Bootstrap messages. In this section, we will consider multi-access subnets where there are both multiple PIM routers in a PIM domain and PIM routers outside the PIM domain or non-trusted hosts. On such subnets, one should (if possible) configure the PMBRs to drop Bootstrap messages. This is possible provided that the routers in the PIM domain receive Bootstrap messages on other internal subnets. That is, for each of the routers on the multi-access subnet that are in our domain, the RPF interface for each of the Candidate-BSR addresses must be an internal interface (an interface not on a multi-access subnet). There are however network topologies where this is not possible. For such topologies, we recommend that IPsec Authentication Header (AH) is used to protect communication between the PIM routers in the domain, and that such routers are configured to drop and log communication attempts from any nodes that do not pass the authentication check. When all the PIM routers are under the same administrative control, this authentication may use a configured shared secret. In order to prevent replay attacks, one will need to have one security association (SA) per sender and use the sender address for SA lookup. The securing of interactions between PIM neighbors is discussed in more detail in the Security Considerations section of [1], and so we do not discuss the details further here. The same security mechanisms that can be used to secure PIM Join, Prune, and Assert messages should also be used to secure Bootstrap messages. How exactly to secure PIM link-local messages is still being worked on by the PIM working group; see [10].

6.4. Candidate-RP-Advertisement Message Security

Even if it is not possible to subvert Bootstrap messages, an attacker might be able to perform most of the same attacks by simply sending C-RP-Adv messages to the BSR specifying the attacker's choice of RPs. Thus, it is necessary to control the sending of C-RP-Adv messages in essentially the same ways that we control Bootstrap messages. However, C-RP-Adv messages are unicast and normally travel multiple hops, so controlling them is more difficult.

6.4.1. Non-Cryptographic Security of C-RP-Adv Messages

We recommend that PMBRs are configured to drop C-RP-Adv messages. One might configure the PMBRs to drop all unicast PIM messages (Bootstrap message, Candidate-RP Advertisement, PIM register, and PIM register stop). PMBRs may also perform source-address filtering to prevent packets entering the PIM domain with IP source addresses that are infrastructure addresses in the PIM domain. We also recommend that implementations have a way of restricting which IP addresses the BSR accepts C-RP-Adv messages from. The BSR can then be configured to only accept C-RP-Adv messages from infrastructure addresses or the subset used for Candidate-RPs.

If the unicast and multicast topologies are known to be congruent, the following checks should be made. On interfaces that are configured to be leaf subnets, all C-RP-Adv messages should be dropped. On multi- access subnets with multiple PIM routers and hosts that are not trusted, the router can at least check that the source Media Access Control (MAC) address is that of a valid PIM neighbor.

6.4.2. Cryptographic Security of C-RP-Adv Messages

For true security, we recommend that all C-RPs are configured to use IPsec authentication. The authentication process for a C-RP-Adv message between a C-RP and the BSR is identical to the authentication process for PIM Register messages between a DR and the relevant RP, except that there will normally be fewer C-RPs in a domain than there are DRs, so key management is a little simpler. We do not describe the details of this process further here, but refer to the Security Considerations section of [1]. Note that the use of cryptographic security for C-RP-Adv messages does not remove the need for the non-cryptographic mechanisms, as explained above.

6.5. Denial of Service using IPsec

An additional concern is that of denial-of-service attacks caused by sending high volumes of Bootstrap messages or C-RP-Adv messages with invalid IPsec authentication information. It is possible that these messages could overwhelm the CPU resources of the recipient.

The non-cryptographic security mechanisms above restrict from where unicast Bootstrap messages and C-RP-Adv messages are accepted. In addition, we recommend that rate-limiting mechanisms can be configured, to be applied on receipt of unicast PIM packets. The rate-limiter MUST independently rate-limit different types of PIM packets -- for example, a flood of C-RP-Adv messages MUST NOT cause a rate limiter to drop low- rate Bootstrap messages. Such a rate-limiter might itself be used to cause a denial-of-service attack by causing valid packets to be dropped, but in practice this is more likely to constrain bad PIM messages. The rate-limiter will prevent attacks on PIM from affecting other activity on the receiving router, such as unicast routing.

7. Contributors

Bill Fenner, Mark Handley, Roger Kermode, and David Thaler have contributed greatly to this document. They were authors of this document up to version 03, and much of the current text comes from version 03.

8. Acknowledgments

PIM-SM was designed over many years by a large group of people, including ideas from Deborah Estrin, Dino Farinacci, Ahmed Helmy, Steve Deering, Van Jacobson, C. Liu, Puneet Sharma, Liming Wei, Tom Pusateri, Tony Ballardie, Scott Brim, Jon Crowcroft, Paul Francis, Joel Halpern, Horst Hodel, Polly Huang, Stephen Ostrowski, Lixia Zhang, Girish Chandranmenon, Pavlin Radoslavov, John Zwiebel, Isidor Kouvelas, and Hugh Holbrook. This BSR specification draws heavily on text from RFC 2362.

Many members of the PIM Working Group have contributed comments and corrections for this document, including Christopher Thomas Brown, Ardas Cilingiroglu, Murthy Esakonu, Venugopal Hemige, Prashant Jhingran, Rishabh Parekh, and Katta Sambasivarao.

9. Normative References

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

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

[3] Meyer, D., "Administratively Scoped IP Multicast", BCP 23, RFC 2365, July 1998.

[4] Deering, S., Haberman, B., Jinmei, T., Nordmark, E., and B. Zill, "IPv6 Scoped Address Architecture", RFC 4007, March 2005.

[5] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

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

10. Informative References

[7] Estrin, D., et al., "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June 1998.

[8] Kim, D., Meyer, D., Kilmer, H., and D. Farinacci, "Anycast Rendevous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)", RFC 3446, January 2003.

[9] Farinacci, D. and Y. Cai, "Anycast-RP Using Protocol Independent Multicast (PIM)", RFC 4610, August 2006.

[10] Atwood, W. and S. Islam, "Security Issues in PIM-SM Link-local Messages", Work in Progress, July 2007.

[11] IANA, "Address Family Numbers", <http://www.iana.org/assignments/address-family-numbers>.

Authors' Addresses

Nidhi Bhaskar
Arastra, Inc.
P.O. Box 10905
Palo Alto, CA 94303
USA
EMail: nidhi@arastra.com

Alexander Gall
SWITCH
P.O. Box
CH-8021 Zurich
Switzerland
EMail: alexander.gall@switch.ch

James Lingard
Arastra, Inc.
P.O. Box 10905
Palo Alto, CA 94303
USA
EMail: jchl@arastra.com

Stig Venaas
UNINETT
NO-7465 Trondheim
Norway
EMail: venaas@uninett.no