Internet Engineering Task Force **PIM WG**
INTERNET-DRAFT
draft-ietf-pim-sm-bsr-00.ps
Bill Fenner/AT&T
Mark Handley/ACIRI
David Thaler/Microsoft
23 February 2001
Expires: August 2001

Bootstrap Router (BSR) Mechanism for PIM Sparse Mode

Status of this Document

This document is a product of the IETF PIM WG. Comments should be addressed to the authors, or the WG's mailing list at pim@catarina.usc.edu.

Abstract

This document specifies the Bootstrap Router (BSR) mechanism for PIM Sparse Mode. BSR is one way that a PIM-SM 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 3
1.1. Overview of Bootstrap and RP Discovery 3
1.2. Administratively Scoped Multicast 4
2. BSR State and Timers 5
3. Bootstrap Router Election and RP-Set Distribution 5
3.1. Sending Candidate-RP-Advertisements 12
3.2. Creating the RP-Set at the BSR 13
3.3. Forwarding Bootstrap Messages 13
3.4. Receiving and Using the RP-Set 14
4. Message Formats 14
4.1. Bootstrap Message Format 15
4.2. Candidate-RP-Advertisement Format 18
5. Default Values for Timers 19
6. Authors' Addresses 20
7. References 20
8. Acknowledgments 20

1. Introduction

For correct operation, every PIM router within a PIM domain must be able to map a particular multicast group address to the same RP. If this is not the case then black holes may appear, where some receivers in the domain cannot receive some groups. A domain in this context 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.

A notable exception to this is where a PIM domain is broken up into multiple administrative scope regions - these are regions where a border has been configured so that a range of multicast groups will not be forwarded across that border. For more information on Administratively Scoped IP Multicast, see RFC 2365. The modified criteria for admin-scoped regions are that the region is convex with respect to forwarding based on the MRIB, and that all PIM routers within the scope region map a particular scoped group to the same RP within that region.

The PIM-SM specification does not mandate the use of a single mechanism to provide routers with the information to perform the group-to-RP mapping. This document describes the Bootstrap Router (BSR) mechanism. BSR was first defined in RFC 2362, which has since been obsoleted. This document provides an updated specification of the BSR mechanism from RFC 2362, and also extends it to cope with administratively scoped region boundaries.

1.1. Overview of Bootstrap and RP Discovery

A small set of routers from a domain are configured as candidate bootstrap routers (C-BSRs) and, through a simple election mechanism, a single BSR is selected for that domain. A set of routers within a domain are also configured as candidate RPs (C-RPs); typically these will be the same routers that are configured as C-BSRs. Candidate RPs periodically unicast Candidate-RP-Advertisement messages (C-RP-Advs) to the BSR of that domain, advertising their willingness to be an RP. A C-RP-Adv message includes the address of the advertising C-RP, as well as an optional list of group addresses and a mask length fields, indicating the group prefix(es) for which the candidacy is advertised. The BSR then includes a set of these Candidate-RPs (the RP-Set), along with their corresponding group prefixes, in Bootstrap messages it periodically originates. Bootstrap messages are distributed hop-by-hop throughout the domain.

All the PIM routers in the domain receive and store Bootstrap messages originated by the BSR. When a DR gets a indication of local membership from IGMP or a data packet from a directly connected host, for a group for which it has no forwarding state, the DR uses a hash function to map the group address to one of the C-RPs from the RP-Set whose group-prefix includes the group (see RFC xxxx). The DR then sends a Join message towards that RP if the local host joined the group, or it Register-encapsulates and unicasts the data packet to the RP if the local host sent a packet to the group.

A Bootstrap message indicates liveness of the RPs included therein. If an RP is included in the message, then it is tagged as ‘up’ at the routers; RPs not included in the message are removed from the list of RPs over which the hash algorithm acts. Each router continues to use the contents of the most recently received Bootstrap message from the BSR until it receives a new Bootstrap message.

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. This time will be on the order of the region's round-trip time and the bootstrap router timeout value.

1.2. Administratively Scoped Multicast

Administratively Scoped IP Multicast, as defined in RFC 2365, permits a network provider to configure scope boundaries at multicast routers. Such a scope boundary consists of a range of multicast addresses (expressed as an address and mask) that the router will not forward across the boundary. For correct operation, such a scope zone border must be complete and convex. By this we mean that there must be no path from inside the scoped zone to outside it 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.

For PIM-SM using BSR to function correctly with admin scoping, there must be a BSR and at least one C-RP within every admin scope region. Admin scope zone boundaries must be configured at the Zone Border Routers (ZBRs), as they need to filter PIM Join messages that might inadvertantly cross the border due to error conditions. However we do not wish any other router within the scope zone to require manual configuration as this creates further possibilities for error, and makes the configuration of large scope zones difficult. If all the C-BSR and C-RP routers within a scope zone are ZBRs, then there is no problem, but this may not the the desired case. Thus whilst we also permit interior C-BSRs and C-RPs to be configured for the admin scope zone, we would also require a mechanism by which all C-BSRs and C-RPs inside an admin scope zone can automatically learn of the existence of the scope zone.

We do this by requiring all ZBRs to be both C-BSRs and C-RPs for the scoped group range, although the default priority should be the lowest possible. A ZBR that does not know of a higher-priority BSR advertising RPs for the scope zone will therefore originate its own Bootstrap message, initially only containing itself as a possible RP for the scoped group range. The group-range field in the ZBRs bootstrap message is marked (using the "Admin Scope" bit, previously a "Reserved" bit) to indicate that this is an administrative scope range, and not just a range that a particular set of RPs are configured to handle. Such a bootstrap message is flooded in the normal way, but will not be forwarded by another ZBR across the boundary for that scope zone (see Section 3.3 for the specifics of this).

When a C-BSR within the scope zone receives such a Bootstrap message, it stores state for the admin scope range contained in the message. A separate BSR election will then take place for every admin scope range (plus one for the global range).

When a C-RP within the scope zone receives such a Bootstrap message, it also stores state for the admin scope range contained in the message. It separately unicasts Candidate-RP-Advertisement messages to the BSR for every admin scope range within which it is willing to serve as an RP. Unless configured otherwise, all candidate RPs are willing to serve as RPs for all groups in all ranges.

ZBRs are also C-RPs for the admin scope zone; they also learn of the current BSR for the admin scope range from receiving a Bootstrap message, and so they must also send a Candidate-RP-Advertisement message to the BSR for the scope range. However, unlike an internal C-RP, a ZBR sets the "Admin Scope" bit in the group-range field in its C-RP advertisement. When the

BSR receives such a C-RP-Adv message, it updates the scope zone keepalive timer; if this timer ever expires the BSR stops being the BSR for that admin scope zone and flushes all state concerned with the scope zone. In this way, if all the ZBRs are configured to no longer be ZBRs, then the BSR will eventually time out the zone.

Note that so long as at least one reachable internal BSR and C-RP is configured within the scope zone to have better-than-minimum priority, then by default the ZBRs themselves will never actually be used as either the BSR or an RP for the scope zone despite being a C-BSR and C-RP.

2. BSR State and Timers

A PIM-SM router implementing BSR holds the following state in addition to the state needed for PIM-SM operation:

At all routers:

List of Active Scope Zones

Per Scope Zone:

Scope-Zone Expiry Timer: SZT(Z)

Bootstrap State:

- Bootstrap Router's IP Address
- BSR Priority
- Bootstrap Timer (BST)
- List of Scope Group-Ranges for this BSR

RP Set

At a Candidate BSR:

Per Scope Zone:

- My state: One of "**Candidate-BSR**", "**Pending-BSR**", "**Elected-BSR**"

At a router that is not a Candidate BSR:

Per Scope Zone:

- My state: One of "**Accept Any**", "**Accept Preferred**"

Bootstrap state is described in section 3, and the RP Set is described in section 3.4.

The following timers are also required:

At the Bootstrap Router only:

Per Scope Zone (Z):

Per Candidate RP (C):

C-RP Expiry Timer: CET(C,Z)

At the C-RPs only:

C-RP Advertisement Timer: CRPT

3. Bootstrap Router Election and RP-Set Distribution

For simplicity, bootstrap messages (BSMs) are used in both the BSR election and the RP-Set distribution mechanisms.

The state-machine for bootstrap messages depends on whether or not a router has been configured to be a Candidate-BSR. The 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.

Candidate-BSR State Machine

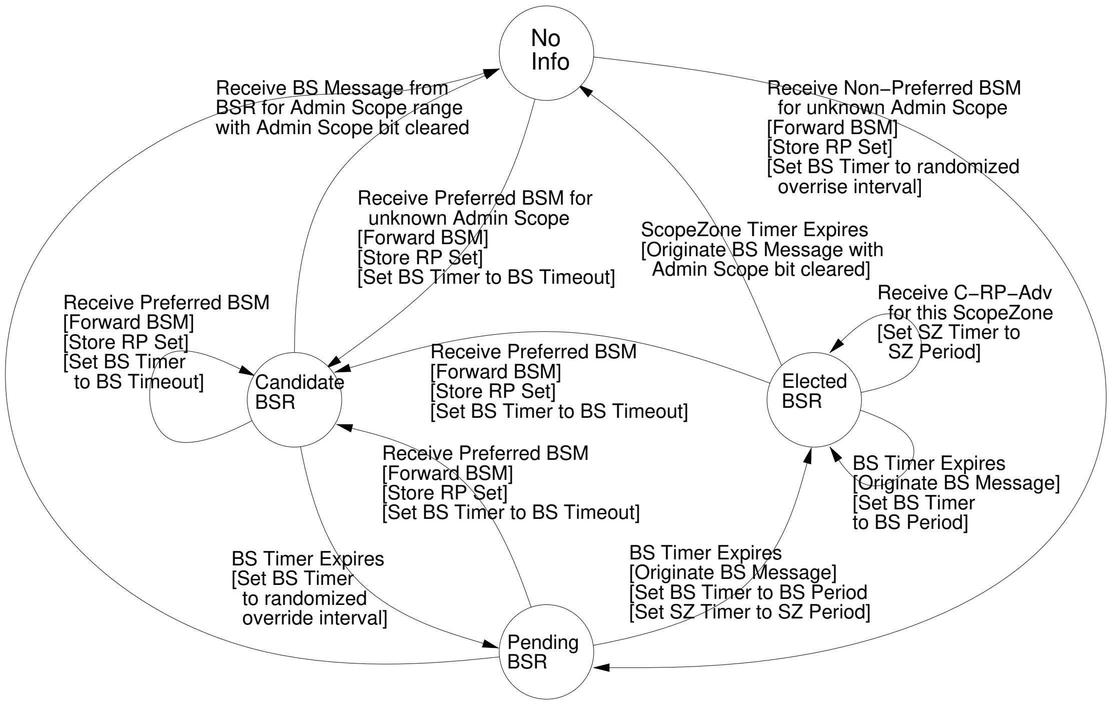

Figure 1: State-machine for a candidate BSR

In tabular form this state machine is:

When in No Info state		
Event	Receive Preferred BSM for unknown Admin Scope	Receive Non-Preferred BSM for unknown Admin Scope
Action	-> C-BSR state Forward BSM; Store RP Set; Set BS Timer to BS Timeout	-> P-BSR state Forward BSM; Store RP Set; Set BS Timer to BS Timeout

When in C-BSR state			
Event	Receive Preferred BSM	BS Timer Expires	Receive BSM from BSR with Admin Scope bit cleared
Action	-> C-BSR state Forward BSM; Store RP Set; Set BS Timer to BS Timeout	-> P-BSR state Set BS Timer to rand_override	-> No Info state cancel timers, clear state

When in P-BSR state			
Event	Receive Preferred BSM	BS Timer Expires	Receive BSM from BSR with Admin Scope bit cleared
Action	-> C-BSR state Forward BSM; Store RP Set; Set BS Timer to BS Timeout	-> E-BSR state Originate BSM; Set BS Timer to BS Period; Set SZ Timer to SZ Period	-> No Info state cancel timers, clear state

When in E-BSR state				
Event	Receive Preferred BSM	BS Timer Expires	SZ Timer Expires	Receive C-RP-Adv for this Admin Scope
Action	-> C-BSR state Forward BSM; Store RP Set; Set BS Timer to BS Timeout	-> E-BSR state Originate BSM; Set BS Timer to BS Period	-> No Info state Originate BSM with Admin Scope bit cleared	-> E-BSR state Set SZ Timer to SZ Timeout

A candidate-BSR may be in one of four states for a particular scope zone:

No Info
The router has no information about this scope zone. This state does not apply if the router is configured to know about this scope zone, or for the global scope zone. When in this state, no state information is held and no timers run that refer to this scope zone.

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

Pending-BSR (P-BSR)
The router is a candidate to be a BSR. Currently no other router is the preferred BSR, but this router is not yet the BSR. For comparisons with incoming BS messages, the router treats itself as the 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 bootstrap router and it must perform all the BSR functions.

On startup, the initial state for this scope zone is "Pending-BSR" for routers that know about this scope zone, either through configuration or because the scope zone is the global scope which always exists; the BS Timer is initialized to the BS Timeout value. For routers that do not know about a particular scope zone, the initial state is No Info; no timers exist for the scope zone.

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

- The bootstrap timer (BS Timer) - that is used to time out old bootstrap router information, and used in the election process to terminate P-BSR state.
- The scope zone timer (SZ Timer) - that is used to time out the scope zone itself at an Elected BSR if no C-RP-Adv messages arrive from the Zone Border Routers.

State-machine for Non-Candidate-BSR Routers

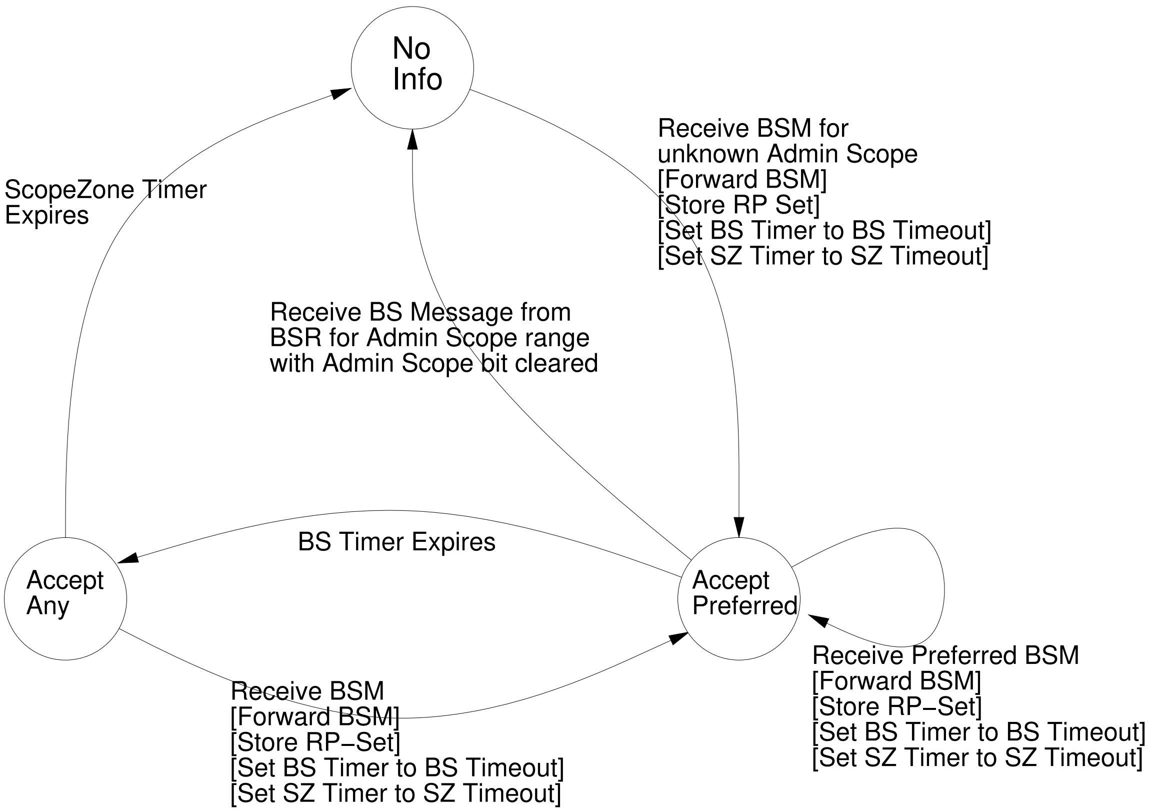

Figure 2: State-machine for a router not configured as C-BSR

In tabular form this state machine is:

When in No Info state	
Event	Receive BSM for unknown Admin Scope
Action	-> AP State Forward BSM; Store RP-Set; Set BS Timer to BS Timeout; Set SZ Timer to SZ Timeout

When in Accept Any state		
Event	Receive BSM	SZ Timer Expires
Action	-> AP State Forward BSM; Store RP-Set; Set BS Timer to BS Timeout	-> No Info state cancel timers; clear state

When in Accept Preferred state			
Event	Receive Preferred BSM	BS Timer Expires	Receive BSM from BSR with Admin Scope bit cleared
Action	-> AP State Forward BSM; Store RP-Set; Set BS Timer to BS Timeout	-> AA State	-> No Info state cancel timers;clear state

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

No Info

The router has no information about this scope zone. This state does not apply if the router is configured to know about this scope zone, or for the global scope zone. When in this state, no state information is held and no timers run that refer to this scope zone.

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. If the router has an RP-Set cached from an obsolete bootstrap message, it continues to use it.

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.

On startup, the initial state for this scope zone is "Accept Any" for routers that know about this scope zone, either through configuration or because the scope zone is the global scope which always exists; the SZ Timer is considered to be always running for such scope zones. For routers that do not know about a particular scope zone, the initial state is No Info; no timers exist for the scope zone.

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

- The bootstrap timer (BS Timer) - that is used to time out old bootstrap router information.
- The scope zone timer (SZ Timer) - that is used to time out the scope zone itself if BS messages specifying this scope zone stop arriving.

Bootstrap Message Processing Checks

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

```
if (BSM.dst_ip_address == ALL-PIM-ROUTERS group) {
  if ( BSM.src_ip_address != RPF_neighbor(BSM.BSR_ip_address) ) {
     drop the BS message silently
  }
} else if (BSM.dst_ip_address is one of my addresses) {
  if ( (No previous BSM received)
       OR (DirectlyConnected(BSM.src_ip_address) == FALSE) ) {
     #the packet was unicast, but this wasn't
     #a quick refresh on startup
     drop the BS message silently
  }
} else {
  drop the BS message silently
}
if (the interface the message arrived on is an Admin Scope
    border for the BSM.first_group_address) {
  drop the BS message silently
}
```

Basically, the packet must have been sent to the ALL-PIM-ROUTERS group by the correct upstream router towards the BSR that originated the BS message, or the router must have no BSR state (it just restarted) and have received the BS message by unicast from a directly connected neighbor. 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 BS message.

BS 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 greater than or equal weight than the current BSR. In a router is in P-BSR state, then it uses its own weight as that of the current BSR.

  The weighting for a BSR is 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).

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

Receive BSM from BSR with Admin Scope bit cleared
: The scope is not the global scope; it is an admin scope that was previously learned from receiving a bootstrap message that had the Admin Scope bit set for this scope. Now a bootstrap message is received for this scope range from the BSR, but the Admin Scope bit is cleared indicating that the BSR has timed out the entire scope zone.

Receive C-RP-Adv for this Admin Scope
: The scope is not the global scope; it is an admin scope range. A C-RP-Adv message arrives with the Admin Scope bit set for this scope range. This indicates that the

sender of the C-RP-Adv (normally a ZBR for the scope zone) believes the scope zone is still active.

BS State-machine Actions

The state-machines specify actions that include setting the BS timer to the following values:

BS Period
The periodic interval with which bootstrap messages are normally sent. The default value is 60 seconds.

BS Timeout
The interval after which bootstrap router state is timed out if no bootstrap message from that router has been heard. The default value is 2.5 times the BS Period, which is 150 seconds.

Randomized Override Interval
The randomized interval during which a router avoids sending a bootstrap message while it waits to see if another router has a higher bootstrap weight. This interval is to reduce control message overhead during BSR election. The following pseudocode is proposed as an efficient implementation of this "randomized" value:

```
Delay = 5 + 2 * log_2(1 + bestPriority - myPriority)
        + AddrDelay
```

where `myPriority` is the Candidate-BSR's configured priority, and `bestPriority` equals:

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

and `AddrDelay` is given by the following:

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

where `myAddr` is the Candidate-BSR's address, and `bestAddr` is the stored BSR's address.

SZ Period
The interval after which a router will time out an Admin Scope zone that it has dynamically learned. The interval MUST be larger than the C-RP-Adv period and the BS Timeout. The default value is ten times the BS Timeout, which is 1500 seconds.

In addition to setting the timer, the following actions may be triggered by state-changes in the state-machines:

Forward BSM
The bootstrap message is forwarded out of all multicast-capable interfaces except the interface it was received on. The source IP address of the message is the forwarding router's IP address on the interface the message is being forwarded from, the destination address is ALL-PIM-ROUTERS, and the TTL of the message is set to 1.

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 multicast-capable interfaces. The IP source address of the message is the forwarding router's IP address on the interface the message is being forwarded from, the destination address is ALL-PIM-ROUTERS, and the TTL of the message is set to 1.

Originate BSM with Admin Scope Bit Cleared
The action is the same as "Originate BSM", except that although this scope zone is an Admin Scope zone, the group range field for the scope zone has the Admin Scope bit cleared. This serves as a signal that the scope zone is no longer in existence.

Store RP Set
The RP-Set from the received bootstrap message is stored and used by the router to decide the RP for each group that the router has state for. Storing this RP Set may cause other state-transitions to occur in the router. The BSR's IP address and priority from the received bootstrap message are also stored to be used to decide if future bootstrap messages are preferred.

In addition to the above state-machine actions, a DR also unicasts a stored copy of the Bootstrap message to each new PIM neighbor, i.e., after the DR receives the neighbor's first Hello message. It does so even if the new neighbor becomes the DR.

3.1. Sending Candidate-RP-Advertisements

Every C-RP periodically unicasts a C-RP-Adv to the BSR for that scope zone to inform the BSR of the C-RP's willingness to function as an RP. Unless configured otherwise, it does this for every Admin Scope zone for which it has state, and for the global scope zone. If the same router is BSR for more than one scope zone, the C-RP-Adv 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 bit MUST be set in the C-RP-Adv messages it sends for that scope zone; otherwise this bit MUST NOT be set.

The interval for sending these messages is subject to local configuration at the C-RP, but must be smaller than the HoldTime in the C-RP-Adv.

A Candidate-RP-Advertisement carries a list of group address and group mask field pairs. This enables the C-RP router to limit the advertisement to certain prefixes or scopes of groups. If the C-RP becomes an RP, it may enforce this scope acceptance when receiving Registers or Join/Prune messages.

The C-RP priority field determines which C-RPs get selected by the BSR to be in the RP Set. Note 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'. ZBRs that do not wish to serve as an RP except under failure conditions should default to sending C-RP-Adv messages with the 'Priority' field set to '255'.

When a C-RP is being shut down, it SHOULD immediately send a C-RP-Adv to the BSR for each scope 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 BSR message removing the shutdown RP from the RPset.

3.2. Creating the RP-Set at the BSR

Upon receiving a C-RP-Adv, if the router is not the elected BSR, it silently ignores the message.

If the router is the BSR, then it adds the RP address to its local pool of candidate RPs. For each C-RP, the BSR holds the following information:

IP address
The IP address of the C-RP.

Group Prefix and Mask list
The list of group prefixes and group masks from the C-RP advertisement.

HoldTime
The HoldTime from the C-RP-Adv message. This is included later in the RP-set information in the Bootstrap Message.

C-RP Expiry Timer
The C-RP-Expiry Timer is used to time out the C-RP when the BSR fails to receive C-RP-Advertisements from it. The expiry timer is initialized to the HoldTime from the RP's C-RP-Adv, and is reset to the HoldTime whenever a C-RP-Adv is received from that C-RP.

C-RP Priority
The C-RP Priority is used to determine the subset of possible RPs to use in the RP-Set.

When the C-RP Expiry Timer expires, the C-RP is removed from the pool of available C-RPs.

The BSR uses the pool of C-RPs to construct the RP-Set which is included in Bootstrap Messages and sent to all the routers in the PIM domain. The BSR may apply a local policy to limit the number of Candidate RPs included in the Bootstrap message. The BSR may override the prefix indicated in a C-RP-Adv unless the 'Priority' field from the C-RP-Adv is less than 128.

The Bootstrap message is subdivided into sets of group-prefix,RP- Count,RP-addresses. For each RP-address, the corresponding HoldTime is included in the "RP-HoldTime" field. The format of the Bootstrap message allows 'semantic fragmentation', if the length of the original Bootstrap message exceeds the packet maximum boundaries. However, we recommend against configuring a large number of routers as C-RPs, to reduce the semantic fragmentation required.

When an elected BSR is being shut down, it should immediately orginate 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.

3.3. Forwarding Bootstrap Messages

Bootstrap messages originate at the BSR, and are forwarded by intermediate routers if they pass the Bootstrap Message Processing Check. Bootstrap messages are multicast to the 'ALL-PIM-ROUTERS' group. When A BS message is forwarded, it is forwarded out of every multicast-capable interface which has PIM neighbors (excluding the one over which the message was received). The exception to this is if the interface is an adminstrative scope boundary for the admin scope zone indicated in the first group address in the BS message packet. Bootstrap messages are always originated or forwarded with an IP TTL value of 1.

3.4. Receiving and Using the RP-Set

When a router receives and stores a new RP-Set, it checks if each of the RPs referred to by existing state (i.e., by (*,G), (*,*,RP), or (S,G,rpt) entries) is in the new RP-Set.

If an RP is not in the new RP-Set, that RP is considered unreachable and the hash algorithm (see PIM-SM specification) is re-performed for each group with locally active state that previously hashed to that RP. This will cause those groups to be distributed among the remaining RPs.

If the new RP-Set contains a RP that was not previously in the RP-Set, the hash value of the new RP is calculated for each group covered by the new C-RP's Group-prefix. Any group for which the new RP's hash value is greater than hash value of the group's previous RP is switched over to the new RP.

4. Message Formats

BSR messages are PIM messages, as defined in RFC xxxx. The values of the PIM message Type field for BSR messages are:

4 Bootstrap Message

8 Candidate-RP-Advertisement

In this section we use the following terms defined in the PIM-SM specification in RFC xxxx:

- Encoded-Unicast format
- 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 [1]. 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 |    Reserved   |Z|  Mask Len     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Group multicast Address
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...
```

Addr Family
described above.

Encoding Type
described above.

Reserved
Transmitted as zero. Ignored upon receipt.

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

Mask Len
The Mask length field is 8 bits. The value is the number of contiguous one bits left justified used as a mask which, combined with the group address, 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 is divided up into ‘semantic fragments’, if the original message exceeds the maximum packet size boundaries. Basically, a single bootstrap message can be sent as multiple packets (semantic fragments), so long as the fragment tage of all the packets comprising the message is the same.

If the bootstrap message contains information about more than one admin scope zone, each different scope zone MUST occupy a different semantic fragment. This allows Zone Border Routers for an admin scope zone to not forward only those fragments that should not traverse the admin scope boundary.

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  |    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 RFC xxxx.

Type
PIM Message Type. Value is 8 for a Bootstrap Message.

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. Note that for historical reasons, the highest BSR priority priority is 255 (the higher the better), whereas the highest RP Priority (see below) is 0 (the lower the better).

Unicast BSR Address
The address of the bootstrap router for the domain. The format for this address is given in the Encoded-Unicast address in RFC xxxx.

Group Address 1..n
The group prefix (address and mask) with which the Candidate RPs are associated. Format described in RFC xxxx. In a fragment containing admin scope ranges, the first group address in the fragment MUST be the group range for the entire admin scope range, and this MUST have the Admin Scope bit set. This is the case 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.

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

RP address 1..m

The address of the Candidate RPs, for the corresponding group prefix. The format for these addresses is given in the Encoded-Unicast address in RFC xxxx.

RP1..m Holdtime

The Holdtime 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 Candidate-RP-Advertisement. 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 per Group Address.

4.2. Candidate-RP-Advertisement Format

Candidate-RP-Advertisements 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 Cnt    |    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 RFC xxxx.

Type

PIM Message Type. Value is 4 for a Candidate-RP-Advertisement Message.

Prefix Cnt

The number of encoded group addresses included in the message; indicating the group prefixes for which the C-RP is advertising. A Prefix Cnt of '0' implies all multicast groups, e.g. for IPv4 a prefix of 224.0.0.0 with mask length of 4. If the C-RP is not configured with Group-prefix information, the C-RP puts a default value of '0' in this field.

Priority

The ‘Priority’ of the included RP, for the corresponding Encoded-Group Address (if any). 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 the advertisement is valid. This field allows advertisements to be aged out.

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 RFC xxxx.

Group Address-1..n

The group prefixes for which the C-RP is advertising. Format described in Encoded-Group-Address in RFC xxxx.

5. Default Values for Timers

Timer Name: Bootstrap Timer (BST)

Value Name	Value	Explanation
BS Period	Default: 60 secs	Period between bootstrap messages
BS Timeout	2 * BS_Period + 10 seconds	Period after last BS message before BSR is timed out and election begins
BS randomized override interval	rand(0, 5.0 secs)	Suppression period in BSR election to prevent thrashing

Timer Name: C-RP Expiry Timer (CET(R))

Value Name	Value	Explanation
C-RP Timeout	from message	Hold time from C-RP-Adv message

C-RP Advertisement messages are sent periodically with period C-RP-Adv-Period. C-RP-Adv-Period defaults to 60 seconds. The holdtime to be specified in a C-RP-Adv message should be set to (2.5 * C-RP-Adv-Period).

Timer Name: C-RP Advertisement Timer (CRPT)

Value Name	Value	Explanation
C-RP-Adv-Period	Default: 60 seconds	Period with which periodic C-RP Advertisements are sent to BSR

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

Value Name Value Explanation		
SZ Timeout	1500 seconds	Interval after which a scope zone will be timed out if the state is not refreshed

6. Authors' Addresses

Bill Fenner
AT&T Labs - Research
75 Willow Road
Menlo Park, CA 94025
fenner@research.att.com

Mark Handley
ACIRI/ICSI
1947 Center St, Suite 600
Berkeley, CA 94708
mjh@aciri.org

David Thaler
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
dthaler@Exchange.Microsoft.com

7. References

[1] IANA, "Address Family Numbers", linked from http://www.iana.org/numbers.html

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, and Pavlin Radoslavov. This BSR specification draws heavily on text from RFC 2362.