Protocol Independent Multicast–Sparse Mode (PIM-SM): Protocol Specification

Stephen Deering
Xerox PARC
3333 Coyote Hill Road
Palo Alto, CA 94304
deering@parc.xerox.com

Deborah Estrin
Computer Science Department/ISI
University of Southern California
Los Angeles, CA 90089
estrin@usc.edu

Dino Farinacci
Cisco Systems Inc.
170 West Tasman Drive,
San Jose, CA 95134
dino@cisco.com

Van Jacobson
Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, CA 94720
van@ee.lbl.gov

Ching-gung Liu, Liming Wei,
Puneet Sharma, and Ahmed Helmy
Computer Science Department
University of Southern California
Los Angeles, CA 90089
charley@catarina.usc.edu, lwei@catarina.usc.edu,
puneet@catarina.usc.edu, ahelmy@catarina.usc.edu

draft-ietf-idmr-PIM-SM-spec-02.ps

September 7, 1995

Status of This Memo

This document is an Internet Draft. Internet Drafts are working documents of the Internet Engineering Task Force (IETF), its Areas, and its Working Groups. (Note that other groups may also distribute working documents as Internet Drafts).

Internet Drafts are draft documents valid for a maximum of six months. Internet Drafts may be updated, replaced, or obsoleted by other documents at any time. It is not appropriate to use Internet Drafts as reference material or to cite them other than as a "working draft" or "work in progress."

Please check the I-D abstract listing contained in each Internet Draft directory to learn the current status of this or any other Internet Draft.

---

[0] Given the length of this author list, it seems appropriate to identify the roles played by each of the authors, who are listed in alphabetical order. Jacobson proposed the original idea of sending join messages towards discovered sources as a means of supporting sparse multicast groups. The detailed architecture and supporting protocols were developed as a collaborative effort of Deering, Estrin, Farinacci, and Jacobson. More recently Liu identified and fixed several critical protocol bugs as part of his implementation effort, and Wei provided data to support the need for shortest path distribution trees (SP-tree) and contributed to protocol development as part of his simulation effort. Puneet Sharma collaborated with Charley Liu on the public-domain implementation of PIM-SM (and the debugging of the spec that coincided with that effort), and most recently Ahmed Helmy helped to integrate the new RP mechanisms and rewrite portions of the spec, including the pseudocode. Estrin, Liu, and Wei were supported by grants from the National Science Foundation and Sun Microsystems.

1 Introduction

This document describes a protocol for efficiently routing to multicast groups that may span wide-area (and inter-domain) internets. We refer to the approach as Protocol Independent Multicast–Sparse Mode (PIM-SM) because it is not dependent on any particular unicast routing protocol, and because it is designed to support sparse groups as defined in [1]. This document describes the protocol details. For the motivation behind the design and a description of the architecture, see [1].

Section 2 summarizes PIM-SM operation. It describes the protocol from a network perspective, in particular, how the participating routers interact to create and maintain the multicast distribution tree. Section 3 describes PIM-SM operations from the perspective of a single router implementing the protocol; this section constitutes the main body of the protocol specification. It is organized according to PIM-SM message type; for each message type we describe its contents, its generation, and its processing. Interoperability with other protocols is discussed in a separate document[2].

Section 4 provides packet format details and section 5 provides pseudocode that corresponds to the functions described in section 3. The pseudocode is just for illustration, Section 4 is authoritative.

The most significant functional changes since the January spec are the RP-related mechanisms (for completeness) and the removal of the PIM-DM protocol details to a separate document[3] (for clarity). We rewrote major portions for clarity and updated the packet formats extensively.

The bibliography, pseudocode, and figures are all in preparation.

2 PIM-SM Protocol Overview

In this section we provide an overview of the architectural components of PIM-SM. PIM-SM protocols operate on group addresses taken from the Sparse portion of the multicast address space.[1]

A PIM-SM router receives explicit join messages from those neighboring routers that have downstream group members. The PIM-SM router then forwards data packets addressed to a multicast group, G, only onto those interfaces on which explicit joins have been received.

A Designated Router (DR) sends PIM-SM Join/Prune messages toward a group-specific Rendezvous Point (RP) for each group for which it has active members. Each router along the path toward the RP builds wildcard (any-source) forwarding state for the group and sends Join/Prune messages on toward the RP. The wildcard forwarding entry's incoming interface points toward the RP; the outgoing interfaces point to the neighboring downstream routers that have sent Join/Prune messages toward the RP. This forwarding state creates a shared, RP-centered, distribution tree that reaches all group members. When a data source first sends to a group its DR unicasts Register messages to the RP with the source's data packets encapsulated within. If the data rate is high, the RP will send source-specific Join/Prune messages back towards the source and the source's data packets will follow the resulting forwarding state and travel unencapsulated to the RP. Whether they arrive encapsulated or natively, the RP forwards the source's decapsulated data packets down the RP-centered distribution tree toward group members. If the data rate warrants it, routers with local receivers can join a source-specific, shortest path, distribution tree, and prune these source's packets off of the shared RP-centered tree. Even if all receivers switch to the shortest path tree, state for that source will be kept at the RP, so that new members that join the RP-centered tree will receive data packets from the source. For low data rate sources, neither the RP, nor last hop routers need join a source-specific shortest path tree and data packets can be delivered via the shared, RP-tree.

[1]Non-SM group addresses may be treated using PIM-DM[3], DVMRP[4], or a locally-configured default RP-list in conjunction with the PIM-SM mechanisms described here. See [2] for more details of the latter.

Figure 1: Example: how a receiver joins, and sets up shared tree
Actions are numbered in the order they occur

The following subsections describe SM operation in more detail, in particular, the control messages that travel up and down the distribution tree, and the actions they trigger. Section 3 describes protocol operation from an implementors perspective, i.e., the actions performed by a single PIM-SM router.

2.1 Local hosts joining a group

In order to join a multicast group, G, a host sends an IGMP Host-Report message identifying the particular group. As specified in [5], IGMP Host-Report messages are sent in response to a directly-connected router's IGMP Host-Query message (see figure 1). From this point on we refer to such a host as a receiver, R, (or member) of the group G. The host also responds with an IGMP RP-Report message identifying the (small) list of RPs associated with the group, referred to as RP-list.[6]

When a DR receives a report for a new group, G, the DR will determine based on the multicast address whether the group is a Sparse Mode group; a specific portion of the multicast address space is being allocated to Sparse Mode groups. If the group is SM the DR looks up the associated RP-list, (see section 2.6), and selects the primary RP. The DR (e.g., router A in figure 1) creates a wildcard multicast forwarding entry for the group, referred to here as a (*,G) entry.

The RP address is included in a special record in the forwarding entry, so that it will be included in upstream Join/Prune messages. The outgoing interface is set to that over which the IGMP Host-Report was received from the new member. The incoming interface is set to the interface used to send unicast packets to the RP. A wildcard bit (WC-bit) associated with this entry is set, indicating that this is a wildcard entry; if there is no more specific match for a particular source, it will be forwarded according to this entry. An RP-bit associated with this entry is also set, indicating that this entry, (*,G), represents state on the shared, RP tree. Each router on the RP-tree with directly connected members sets a timer for this entry. The RP-timer is reset each time an RP-Reachability message is received for (*,G), see section 2.2. [2] If the timer expires and the router has no local members, the (*,G) state is deleted. If the

[2] Optionally, a router without directly connected members may also process RP-reachability messages and thereby timeout (*,G) state more rapidly. However, this is not required for proper function of the protocol since the routers with directly connected members will eventually time out their entries and stop sending (*,G) Join/Prune messages toward the unreachable RP.

router does have local members, it refreshes the (*,G) entry timer each time it gets an IGMP membership report; then, when the RP-timer expires, the router attempts to join to the next RP on the RP-list.

2.2 Establishing the RP-rooted shared tree

Triggered by the (*,G) state, the DR creates a Join/Prune message with the RP address in its join list and the WC-bit and RP-bit set; nothing is listed in its prune list. The RP-bit flags the join as being associated with the shared tree and therefore the Join/Prune message is propagated along the RP-tree. The WC-bit indicates that the address is an RP and the receiver expects to receive packets from all sources via this (shared tree) path.

Each upstream router creates or updates its multicast forwarding entry for (*,G) when it receives a Join/Prune with the RP-bit and WC-bit set. The interface on which the Join/Prune message arrived is added to the list of outgoing interfaces (oifs) for (*,G). Based on this entry each upstream router between the receiver and the RP sends a PIM-SM- Join/Prune message in which the join list includes the RP. The packet payload contains Multicast-Address=G, Join={RP,WCbit,RPbit}, Prune=NULL.

When a router that is willing to act as an RP receives a (*,G) Join with itself listed as the RP, the router automatically performs the functions specified here for an RP (i.e., a router does not need to be specially configured to act as an RP). The incoming interface (iif) in the RP's (*,G) entry is set to null. RP-Reachability messages are generated by the RP periodically and distributed down the (*,G) tree established for the group. This allows downstream routers to detect when their current RP has become unreachable and trigger joining towards an alternate RP, see section 2.6. When alternate RPs are used, (*,G) Join/Prune messages include the complete ordered RP-list. An RP performs the functions described thus far whether it is the primary RP, or an alternate; however, alternate RPs have the added task of polling preferred RPs on the RP-list and notifying leaf routers when a preferred RP becomes reachable.

2.3 Hosts sending to a group

To start sending packets to a group, a host responds to IGMP queries from the DR with an IGMP RP-Report for that group; the IGMP message is sent to the "RP-Reporters" group with a TTL of 1. All PIM hosts on the LAN join this group in order to implement suppression (see figure 2). The DR stores the indicated RP-Group mapping. When a host first sends a multicast data packet to a group, its DR must deliver the packet to the RP for distribution down the RP-tree. This is done by the sender's DR unicasting a PIM-SM-Register packet to the primary RP for the group (see section 2.6). The data packet is encapsulated in the PIM-SM-Register packet so that the RP can decapsulate it and deliver it to downstream members. The RP responds to Registers with explicit Register-Ack messages. These Register-Ack messages are sent periodically, and provide liveness indication; their absence does not trigger retransmission, it triggers the router to select an alternate RP.

If the data rate of the source warrants the use of a source-specific shortest path tree, the RP constructs (S,G) state and sends periodic Join/Prune messages toward the source. The routers between the source and the RP build and maintain (S,G) state in response to these messages and send (S,G) messages upstream toward the source. Each (S,G) state entry includes the RP-address in the RP-Annotated field of the entry. S,G Join/Prune messages triggered off of that state will include the RP-address.

The source's DR stops encapsulating data packets in PIM-SM-Registers when (and so long as) it receives Join/Prune(S,G) messages from the active RP (i.e., S's RP Annotated-bit is set in the join list). Even if the RP does not set up (S,G) state, it still responds to the source's Register messages with Register-Acks, when requested (i.e., if the Ack-Request flag is set in the Register message). If the RP has a (S,G,RPbit) entry or (*,G) entry with a null oif list, the RP sets a no-data flag in the Register-Ack

Figure 2: Example: a host sending to a group
Actions are numbered in the order they occur

to suppress the source's DR from encapsulating the sources data packets. If the RP has no state at all for that group, it responds with no-data Register-Acks. To deal with a failure scenario in which the primary RP is unreachable for extended periods and data sources are very bursty, the DR continues to send null-Register messages periodically so long as directly connected sources continue to send IGMP RP-reports; this is only necessary when the active RP is not the primary RP.

If an RP has gone down during the register process, we want to limit how long we encapsulate data packets. Also, if the encapsulating stops and data is forwarded via (S,G) state to the RP, it is desirable to know if that RP becomes unreachable. Therefore, there is an RP (liveness) timer, and an RP-status flag, kept for the active RP for all active groups in the DR of each source. The RP-timer is reset, and the RP-status flag is set to "reachable" when a Join/Prune with the RP address included, or a Register-Ack message, is received. When the RP-timer expires (for example, 270 seconds), the RP-status flag is set for that RP indicating that it is in a "down" state. The actions taken when an RP is detected as being unreachable are described in section2.6.

2.4 Switching from shared tree (RP-tree) to shortest path tree (SP-tree)

When a PIM router has directly-connected members it first joins the shared RP-tree. The router can switch to a source's shortest path tree (SP-tree) after receiving packets from that source over the shared RP-tree. The recommended policy is to initiate the switch to the SP-tree after receiving a significant number of data packets during a specified time interval from a particular source. To realize this policy the router monitors data packets from sources for which it has *no* source-specific multicast forwarding entry and initiates such an entry when the data rate exceeds the configured threshold. As shown in figure 3, router A initiates a new multicast forwarding entry that is specific to the source, hereafter referred to as (S,G) state.

When a (S,G) entry is activated (and periodically so long as the state exists), a Join/Prune message will be sent upstream towards the source, S, with S in the join list. The payload contains Multicast-Address=G, Join={S}, Prune=NULL. When the (S,G) entry is created, the outgoing interface list is copied from (*,G), i.e., all local shared tree branches are replicated in the new shortest path tree. In this way when a data packet from S arrives and matches on this entry, all receivers will continue to receive

Figure 3: Example: Switching from shared tree to shortest path tree
Actions are numbered in the order they occur

the source's packets along this path unless and until the receivers choose to prune themselves. Note that (S,G) state must be maintained in all last-hop routers where an SP-tree is maintained. Even when (*,G) and (S,G) overlap, both states are needed to trigger the source-specific Join/Prune messages. (S,G) state is kept alive by data packets arriving from that source. A timer, S-timer, is set for the (S,G) entry and this timer is reset whenever a data packet for (S,G) is received. When the S-timer expires the state is deleted.

Only routers with local members can initiate switching to the SP-tree; intermediate routers do not. Consequently last hop routers initialize (S,G) state in response to data packets from the source, S; whereas intermediate routers only initialize (S,G) state in response to Join messages from downstream that have S in the Join list. To implement the policy that source-specific trees are only setup for high-data rate source, a last-hop router does not initialize a (S,G) entry until it has received m data packets from the source within some interval of n seconds. The last-hop router may alternatively be configured to not request switching to the shortest path tree.

The (S,G) entry is initialized with the SPT-bit cleared, indicating that the shortest path tree branch from S has not been setup completely, and the router can still accept packets from S that arrive on the (*,G) entry's iif. When a router with a (S,G) entry and a cleared SPT-bit starts to receive packets from the new source S on the iif for the (S,G) entry, and that iif differs from the (*,G) entry's iif, the router sets the SPT-bit, and sends a Join/Prune message towards the RP. This indicates that the router no longer wants to receive packets from S via the shared RP-tree. The Join/Prune message sent towards the RP includes S in the prune list, with the RP-bit set indicating that S's packets should not be forwarded down this branch of the shared tree. If the router receiving the Join/Prune message has (S,G) state (with or without the RPbit set), it deletes the arriving interface from the (S,G) oif list. If the router has only (*,G) state, it creates an (S,G,RPbit) entry. The Join/Prune message payload contains Multicast-Address=G, Join=NULL, Prune={S,RPbit}.

If at a later time a new receiver joins the RP-tree, the negative cache state on the RP-tree must be eradicated to bring all sources' data packets down to the new receiver. Therefore, when a (*,G) Join arrives with a null prune list at a router that has any (S,G,RP-bit) entries (which is causing it to send source-specific prunes toward the RP), all RP-bit state for that group has to be updated upstream of the router; so as to bring all sources' packets down to the new member. To accomplish this the router updates all existing (S,G,RP-bit) entries; it adds to each (S,G,RPbit) entry's oif list the interface on which the (*,G) join arrived. The router also triggers a (*,G) join upstream to cause the same updating of RP-bit settings upstream and pull down all active sources' packets. If the arriving (*,G) join has some sources included in its prune list, then the corresponding (S,G,RP-bit) entries are left unchanged (i.e., the RPbit remains set and no oif is added).

2.5 Steady state maintenance of distribution tree (i.e., router state)

In the steady state each router sends periodic Join/Prune messages for each active (S,G) or (*,G) entry; the Join/Prune messages are sent to the RPF neighbor on the iif of the corresponding entry. These messages are sent periodically to capture state, topology, and membership changes. A Join/Prune message is also sent on an event-triggered basis each time a new forwarding entry is established for some new source (note that some damping function may be applied, e.g., a merge time). Join/Prune messages do not elicit any form of explicit acknowledgment; routers recover from lost packets using the periodic refresh mechanism.

2.6 Use of alternate RPs (i.e., Adaptation to RP unreachability)

For each multicast group, the group initiator selects a primary RP and a small ordered set of alternate RPs; referred to as the RP-list.

Except for transients while adapting to failures and recoveries, only a single RP is active per group at any point in time. A later section, 2.7, describes a mechanism to assist group initiators in selecting routers for the RP-list. This section describes the use of the RP-list once it has been constructed and advertised; in particular, the use of alternate RPs when the primary RP becomes unreachable.

When a router receives (*,G) Joins indicating itself as the RP, it sets up (*,G) state and periodically sends RP-reachability messages. These messages traverse the shared RP tree down to last hop routers who use it to reset the timers on their (*,G) state. If a DR's (*,G) state timer expires, this indicates that RP-reachability messages are no longer being received. This triggers the DR to send (*,G) joins toward the next RP on the ordered RP list for that group. When the primary RP becomes unreachable, all DRs on the shared distribution tree will detect the event and switch to the same alternate RP. Consequently, aside from transients, there is always a single shared RP-tree with a single active RP at any point in time. The only time that different receiver's DRs will take different action from one another is when there is a network partition and some DRs still receive reachability messages while others do not. In this case only the receivers on the other side of the partition will initiate joins toward the secondary RP. The (*,G) join sent toward the alternate RP includes the complete ordered list of RPs for that group (for reasons explained below).

If and when the RP becomes unreachable, sources' first hop routers will stop receiving the RP's (S,G) Join or Register-Ack messages. Consequently, the RP timers in the sources' first hop routers will also expire. This will trigger these routers to send subsequent (S,G) data packets encapsulated in Register messages to the next RP in the ordered list. The Register messages include the ordered RP-list.

Since all new members will be joining to the preferred RP once it is reachable, the senders and receivers switch back to the preferred RP when it becomes reachable. To achieve this, an active, alternate RP periodically polls the preferred RPs (all RPs that appear before it in the ordered RP list). [3] When the active alternate RP finds that one of the preferred RPs is reachable, the active RP multicasts a RP-reachability message down the (*,G) tree indicating which RP the last hop DRs should join. It also unicasts a Register-Ack message to the sources' first hop routers informing them of the now-reachable

[3] RPn polls RPn-i, where i=n-1,...,1, so long as RPn is active and and until an RPi responds.

and preferred RP address. A Register-Ack with the preferred RP-address included is sent in response to a sampling of subsequent Register packets received.

The alternate RP may prune any upstream (S,G) state or just allow it to time out. Note that switching back is unlikely to impose significant degradation in performance, since for high data rate sources, receivers will be joined to the SP-tree, and data delivery will not be affected by the switch.

Note that we do not try to fix the case where a receiver can reach the alternate RP and the alternate RP can reach the primary, but the receiver can not reach the primary. This situation could result from inconsistent unicast routing or perhaps an asymmetry caused by a firewall. The former case should be addressed by the unicast routing protocol (and is being so addressed) , and the latter case requires that we articulate to firewall users how their firewalls and PIM-SM routers need to be configured in order to allow PIM usage where desired.

2.7 RP Selection

The mechanism proposed here is one possible means of selecting RPs; it does not preclude the use of alternate methods, heuristics, and even out of band procedures for selecting RPs, so long as the selected RPs are placed in an ordered list and advertised to all potential group members and sources to groups. However, the particular mechanism proposed here will produce more scalable, robust, and efficient RP distribution trees and therefore is important to the overall architecture.

To summarize our approach, we provide a mechanism for the Primary RP to be selected from among routers close to the group initiator, and alternate RPs from other parts of the network, depending upon the anticipated geographic scope of the group. We assume that in general the network is not partitioned and the primary RP is used. Network topology changes will be reflected in routing protocol adaptations and consequent adaptation of the affected branches of the RP (and source specific) tree. Only when the primary RP fails or when the network partitions (i.e., a failure occurs that routing cannot heal), does the protocol automatically switch to one of the alternate RPs specified for a group. In other words, the adaptation mechanism occurs in response to relatively rare events.

Routers that are willing to act as RPs use a low-frequency advertisement protocol as follows:

1. Candidate-RP-Advertisement messages are sent to a well-known, multicast group such as that used by sd for session advertisements.

2. Each message includes an Intended-Hop-Count value set by the advertising router; this value is *not* modified by the other routers which forward the packet to the well-known distribution group. The advertising router initializes the TTL in the containing IP packet to this Intended-hop-count value as a means of controlling the range of its advertisements and its resulting use as an RP. Candidate-RP-Advertisements also include a group address and group mask fields, which convey information about the range of groups for which the advertising router is willing to become an RP. [4]

Hosts that are used for multicast group initiation (e.g., those that now run the sd protocol, or a smaller set of servers that are queried by such hosts) join the Candidate-RP-Advertisement group and receive advertisements from all candidate RP routers whose scope extends far enough. These hosts/servers classify the received advertisements according to the "distance" of the advertising router. The distance of an advertising candidate can be computed based on the advertisement message by subtracting the IP header TTL value from the Intended-hop-count value. For example, in the context of a particular server/host contacted by the group initiator, the local Candidate-RPs might consist of only the current DR or a set of routers and Border Routers in the same domain as the initiator; whereas the regional

[4] If a router has multiple interfaces and is sending candidate RP advertisements, it should advertise its most generally reachable address.

Candidate-RPs might be all those that are within a small number of hops beyond the local domain. Candidate-RP-Advertisements are slowly aged to allow for changes in the candidacy of an RP.

When a group initiator defines a multicast group, it will specify the likely-group-scope. The RP selection tool will then select the primary RP from the local RP-candidate list. The alternate RP list will be constructed by selecting one (possibly 2) RP from each of the candidate list sets that is within the group scope.

Once the alternate RPs have been selected they are placed in an ordered list, with the primary RP first. We assume the existence of an sd-like tool for RP-list advertisement. RP-reports are sent by group participants (receivers and senders) to their directly connected DRs, to inform them of the RP-list.

2.8 Multicast data packet processing

Data packets are processed in a manner similar to existing multicast schemes. A router first performs a longest match on the source and group address in the data packet. A (S,G) entry will be matched first if one exists; a (*,G) entry will be matched otherwise. If neither state exists, then the packet is dropped. An incoming interface check (RPF check) is performed on the matching state and if it fails the packet is dropped, otherwise the packet is forwarded to all interfaces listed in the outgoing interface list.

The following two actions must be introduced in order to deliver data packets continuously during the transition from a shared to shortest path tree. First, when a data packet matches on a (S,G) entry with a cleared SPT-bit, if the packet does not match the incoming interface for that (S,G) entry, but the packet does match the incoming interface for the (*,G) entry, then the packet is forwarded according to the (S,G) entry. In addition, when a data packet matches on a (S,G) entry with a cleared SPT-bit, and the incoming interface of the packet matches that of the (S,G) entry, then the packet is forwarded and the SPT-bit is set for that entry.

Data packets to SM groups never trigger prunes. However, data packets may trigger actions which in turn trigger prunes. For example, when router *B* in figure 3 decides to switch to SP-tree at step 3, it creates a (S,G) entry with SPT-bit set to 0. When data packets from S arrive at interface 2 of *B*, *B* sets the SPT-bit to 1, which in turn triggers the sending of prunes towards the RP.

2.9 Operation over Multi-access Networks

This section describes a few additional protocol mechanisms needed to operate PIM over multi-access networks: Designated Router election, Using Assert messages to resolve parallel paths, and suppressing redundant Joins and Registers on multi-access networks.

2.9.1 Designated router election

When there are multiple PIM routers connected to a multi-access network, one of them should be chosen to operate as the designated router (DR) at any point in time. The DR is responsible for sending IGMP Host-Query messages to solicit host group membership IGMP Host-Reports; the DR is also responsible for initiating (*,G) state to trigger Join/Prune messages toward the RP and keep track of the active RP status for local senders.

A simple designated router (DR) election mechanism is used for both SM and traditional IP multicast routing.

Neighboring routers send PIM-Query packets to each other. The sender with the largest IP address assumes the role of DR. Each PIM router connected to the multi-access LAN sends the PIM-Queries periodically in order to adapt to changes in router status.

2.9.2 Parallel paths to a source or the RP

If a router receives a multicast datagram on a multi-access LAN from a source whose corresponding (S,G) outgoing interface list includes the received interface, the packet must be a duplicate. In this case a single forwarder must be elected. Using PIM-Assert messages addressed to 224.0.0.2 (all routers) on the LAN, upstream routers can decide which one becomes the forwarder. Downstream routers listen to the asserts so they know which one was elected (i.e. typically this is the same as the downstream router's RPF neighbor but there are circumstances when using different unicast protocols where this might not be the case), and therefore where to send subsequent Joins.

The upstream router elected is the one that has the shortest distance to the source. Therefore, when a packet is received on an outgoing interface a router will send a PIM-Assert packet on the multi-access LAN indicating what metric it uses to reach the source of the data packet. The router with the smallest numerical metric will become the forwarder. All other upstream routers will delete the interface from their outgoing interface list. The downstream routers also do the comparison in case the forwarder is different than the RPF neighbor. [5]

Associated with the metric is a metric preference value. This is provided to deal with the case where the upstream routers may run different unicast routing protocols. The numerically smaller metric preference is always preferred. The metric preference should be treated as the high-order part of an assert metric comparison. Therefore, a metric value can be compared with another metric value provided both metric preferences are the same. A metric preference can be assigned per unicast routing protocol and needs to be consistent for all routers on the multi-access network.

Asserts are also needed for (*,G) entries since there may be parallel paths from the RP and sources to a multi-access network. When an assert is sent for a (*,G) entry, the first bit in the metric preference (RP-bit) is always set to 1 to indicate that this path corresponds to the RP tree. Furthermore, the RP-bit is always cleared for SP-tree entries metric preference, this causes an SP-tree path to always look better than an RP-tree path. When the SP-tree and RPtree cross the same LAN, this mechanism eliminates the duplicates that would otherwise be carried over the LAN.

The DR may lose to another router on the LAN by the Assert process if there are multiple paths to the active RP through the LAN. From then on, the DR is no longer the last-hop router for local receivers. The winning router becomes the last-hop router and is responsible for sending (*,G) join messages to the RP. Asserts are rate limited.

2.9.3 Join/Prune suppression

If a Join/Prune message arrives on the incoming interface for an existing (S,G) entry, and the sender of the Join/Prune has a higher IP address than the recipient of the message, a Joiner-bit is cleared to suppress further Join/Prune messages. A timer is set for the Joiner-bit; after it expires the Joiner-bit is set indicating further periodic Join/Prunes should be sent for this entry. The Joiner-bit timer is reset each time a Join/Prune message is received from a higher-IP-addressed PIM neighbor.

2.9.4 Register suppression and Register-Acks

When a router receives a (S,G) join for a source, S, that is directly connected to the router via a multiaccess network, the router must send the join to 0.0.0.0 on the mutliaccess network, in case it is not the DR. This address is used when the upstream router is not known and so the target for the Join/Prune is not known. When a DR receives the Join/Prune on its incoming interface for a directly connected source

[5] The downstream routers will change their upstream neighbor to the router that sent the last PIM-Assert message during the assert process. This is important so downstream routers send subsequent PIM-Joins/Prunes (in SM) to the correct neighbor.

whose RP Annotated-bit is set in the join list, the DR sets its Register timer to suppress the sending of registers for that source. If such Join/Prune messages stop arriving at the DR, its RP register timer will eventually expire and subsequent packets from the source will cause registers to be sent to the RP.

2.10 Unicast Routing Changes

When unicast routing changes, an RPF check is done on all active (S,G) and (*,G) entries, and all affected expected incoming interfaces are updated. In particular, if the new incoming interface appears in the outgoing interface list, it is deleted from the outgoing interface list. The previous incoming interface may be added to the outgoing interface list by a subsequent Join/Prune from downstream. Joins received on the current incoming interface are ignored. Joins received on new interfaces or existing outgoing interfaces are not ignored. Other outgoing interfaces are left as is until they are explicitly pruned by downstream routers or are timed out due to lack of appropriate Join/Prune messages.

The PIM router must send a Join/Prune message with S in the Join list out its new incoming interface to inform upstream routers that it expects multicast datagrams over the interface. It may also send a Join/Prune message with S in the Prune list out the old incoming interface, if the link is operational, to inform upstream routers that this part of the distribution tree is going away.

2.11 Interaction with non-PIM-SM protocols

Interaction with non-PIM-SM networks is discussed in a separate interoperability document[2] [6].

All special mechanisms that deal with interoperability are executed in Border Routers of the PIM-SM region and do not require any modification of regular PIM-SM routers.

2.12 Treatment of non-SM groups

PIM-SM routers may be configured to run a DM protocol to handle non-SM groups, e.g., PIM-DM, DVMRP, or MOSPF[7].

Alternatively, PIM-SM routers may be configured with a default RP-list for use with all non-PIM-SM groups. For SM groups, PIM-SM relies on a group-specific RP-lists that are advertised and used by all members and sources, internet-wide. For non-SM groups, PIM-SM would use a local domain-specific RP-list that is configured and used for all groups, but only within that domain. Each domain would create and configure its own local RP-list. Apart from the local definition of the RP-list, all other PIM-SM mechanisms remain unchanged.

Unlike the other alternatives, this would create a single shared tree within the domain for use by all non-SM groups. PIM-DM, DVMRP, and MOSPF all create source-specific trees.

[6] This document is currently in preparation.

3 Detailed Protocol Description

This section describes the protocol operations from the perspective of an individual PIM router implementation. In particular, for each message type we describe how it is generated and processed. In this version of the spec we suggest particular numerical timer settings. A future version of the spec will specify a mechanism for timers to be set as a function of the outgoing link bandwidth.

3.1 Query

PIM-Query messages are sent so neighboring PIM routers can discover each other.

3.1.1 Sending Queries

Query messages are sent periodically between PIM neighbors. By default they are transmitted every 30 seconds. This informs routers what interfaces have PIM neighbors. Query messages are multicast using address 224.0.0.2. The packet includes the holdtime for neighbors to keep the information valid. The recommended holdtime is 3 times the query transmission interval. By default the holdtime is 90 seconds. Queries are sent on all types of communication links.

3.1.2 Receiving queries

When a router receives a PIM-Query packet, it stores the IP address for that neighbor, sets the PIM neighbor timer based on the PIM-Query holdtime, and determines the Designated Router (DR) for that interface. The highest IP addressed system is elected DR. Each query received causes the DR's address to be updated.

3.1.3 Timing out neighbor entries

A periodic process is run to time out PIM neighbors that have not sent queries. If the DR has gone down, a new DR is chosen by scanning all neighbors on the interface and selecting the new DR to be the one with the highest IP address. If an interface has gone down, the router may optionally time out all PIM neighbors associated with the interface.

3.2 IGMP RP-Reports

Editors Note: This section will be detailed in the next I-D release. We decided at the last moment that although RP-Reports are a part of IGMP and not PIM, per se, that we need the detailed specification of their handling included in the PIM-SM specification. This subsection on IGMP RP-Reports is just a draft and has not been reviewed.

Hosts respond to IGMP-Queries with IGMP RP-Reports if they have live RP-Group mapping information. The RP-Report contains the following information:

- The Group address.
- The ordered list of RP addresses for the group.
- A two bit flag. The receiver-flag bit is set if the host wishes to join the group. The source-flag bit is set if the host intends to send data to the group. At least one bit will be set; both may be set.

The RP-Reports are sent to the RP-Reporters group with a TTL of 1. In addition to the routers that support PIM, all hosts on the LAN that send IGMP RP-Reports join the RP-Reporters group. RP-Report information is suppressed by equivalent information in other recent RP-Reports. Information is equivalent if the Group address and RPlist are the same, and the corresponding Sender/Receiver flags are set.

When a DR receives an IGMP RP-Report message it processes it as follows.

- If no corresponding RP-Group-mapping exists, the DR initializes one. If there exists RP-Group-mapping the RPlist is updated.
- Sets the Source-flag and Receiver-flag bits in the RP-group-mapping state according to the flag setting in the RP-Report.
- Resets the RP-Group-mapping timer associated with the RP-Group-mapping state.
- If the Source flag is set to 1 and the Ack-Request timer for this group is non-existent or has a zero value, then the group's Ack-Request timer is initialized and the Ack-Request flag is set to 1.
- If the Receiver flag is set to 1, the (*,G) state is refreshed or initialized.

3.3 Join/Prune

Join/Prune messages are sent to join or prune a branch off of the multicast distribution tree. A single message contains both a join and prune list, either one of which may be null. Each list contains a set of source addresses, indicating the source-specific trees or shared tree that the router wants to join or prune.

3.3.1 Sending Join/Prune Messages

Join/Prune messages are merged such that a message sent to a particular upstream neighbor, N, includes all of the current joined and pruned sources that are reached via N; according to unicast routing Join/Prune messages are multicast to all routers on multi-access networks with the target address set to the next hop router towards S or RP. Join/Prune messages are sent periodically. Currently the period is set to 60 seconds.[7]

A router will send a periodic Join/Prune message to each distinct RPF neighbor associated with each (S,G) and (*,G) entry.

Join/Prune messages are only sent if the RPF neighbor is a PIM neighbor. A periodic Join/Prune message sent towards a particular RPF neighbor is constructed as follows:

1. The RP address (with RP and WC bits set) is included in the join list, and the RP-list is included in the RP-Address fields, of a periodic Join/Prune message under the following conditions:
   (a) The Join/Prune message is being sent to the RPF neighbor to the RP.
   (b) The active RP is determined to be in Up state, and
   (c) The outgoing interface list in the (*,G) entry is non-NULL, or the router is the DR on the same interface as the RPF neighbor.
2. A particular source address, S, is included in the join list with the RP and WC bits cleared under the following conditions:

[7]In the future we will introduce mechanisms to rate-limit this control traffic on a hop by hop basis, in order to avoid excessive overhead on small links.

(a) The Join/Prune message is being sent to the RPF neighbor to S, and

(b) There exists an active (S,G) entry with the RPbit cleared, and

(c) The *oif* list in the (S,G) entry is not null.

The RP Annotated-bit (A-bit) is set for source S in the join list if the local (S,G) entry has a valid IP address listed in its RP-Annotated field. The (S,G) entry's RP-Annotated field is included in the group's RP-Address-1 field and the RP count is set to 1.

3. A particular source address, S, is included in the prune list with the RP and WC bits cleared (and A-bit cleared) under the following conditions:

   (a) The Join/Prune message is being sent to the RPF neighbor to S, and

   (b) There exists an active (S,G) entry with the RPbit cleared, and

   (c) The *oif* list in the (S,G) entry is null.

4. A particular source address, S, is included in the prune list with the RP bit *set* and the WC bit cleared (and A-bit cleared) under the following conditions:

   (a) The Join/Prune message is being sent to the RPF neighbor toward the RP and there exists a (S,G) entry with the RPbit set and null *oif* list, or

   (b) The Join/Prune message is being sent to the RPF neighbor toward the RP, there exists a (S,G) entry with the RPbit cleared and SPT-bit set, and the incoming interface toward S is different than the incoming interface toward the RP.

In addition to these periodic messages, the following events will trigger Join/Prune messages (the contents of triggered messages are the same as the periodic, described above)

1. Receipt of an IGMP Host-Report message for a new SM group G (i.e., an SM group for which the receiving router does not have a (*,G) entry) will trigger creation of a (*,G) entry and sending of a Join/Prune message towards the RP with the RP address and RP-bit and WC-bits set in the join list.

2. Receipt of a Join/Prune message for (S,G) or (*,G) will cause building or modifying corresponding state, and subsequent triggering of upstream Join/Prune messages, in the following cases:

   (a) When there is no current forwarding entry, an entry will be created. The new entry will in turn trigger an upstream Join/Prune message.

   (b) When the outgoing interface list of (S,G,RPbit) entry is null, the triggered Join/Prune message will contain S in the prune list.

   (c) When a source, S, in the Join/Prune message has its RP Annotated-bit set to zero, and the existing (S,G) entry has the RP-Annotated field set to a valid IP address (the RP's address).

   (d) When the source, S, in the Join/Prune message has its RP Annotated-bit set to one, and the existing (S,G) entry has the RP-Annotated field set to all zeros: the (S,G) entry is updated to correspond to the arriving Join/Prune message and the triggered Join/Prune message reflects the new setting in the entry.

   (e) When an oif times out for which the A-bit was set, and no other oif has the A-bit set, the entry's A-bit and RP-Annotate fields are cleared and a Join/Prune message is triggered upstream to represent the new state status.

3. Receipt of a packet on a (S,G) entry whose SPT-bit is cleared triggers the following if the packet arrived on the correct incoming interface and there is a (*,G) entry with a different incoming RPF neighbor: a) setting of the SPT-bit on (S,G) entry, and b) sending a Prune message towards the RP with S,RP-bit in the prune list if the iif(S,G) does not equal the iif(*,G).

4. When a Join/Prune message is received for a group G, the prune list is checked. If it contains a source for which the receiving router has an active (S,G) entry, and whose *iif* is that on which the Join/Prune was received, then a join for (S,G) is triggered to override the prune. (This is necessary in the case of parallel downstream routers connected to a multi-access network.)

5. When a router receives a Join/Prune message with a source in the join list that is directly connected to the router via a multi-access LAN, the router must send the Join/Prune to 0.0.0.0 on the LAN in case it is not the DR.

6. When the active RP fails, RP-Reachability messages will not reach the receivers' last-hop routers, hence, the (*,G) state RP-timers will expire. This triggers the last-hop routers to send (*,G) joins towards the next RP on the RP list for that group. The Join/Prune message to the alternate RP includes the ordered RP-list.

7. When an active alternate RP finds one of the preferred RPs reachable, the active RP sends a special RP-reachability message down the (*,G) tree indicating to which RP the last-hop routers should join. This triggers updating of the (*,G) state at the last hop routers, which in turn triggers sending of a (*,G) Join upstream.

We do not trigger prunes onto interfaces for SM groups based on data packets. Data packets that arrive on the wrong incoming interface for an SM group are silently dropped.

3.3.2 Receiving Join/Prune Messages

When a router receives a Join/Prune message, it processes it as follows:

1. The receiver of the Join/Prune notes the interface on which the PIM message arrived, call it I. The router accepts this Join/Prune message if this Join/Prune message is addressed to the router itself. If the Join/Prune is for this router the following actions are taken:

   (a) If an address Sj in the join list has RP-bit and WC- bit set, then Sj is an RP address and the following actions are taken:

      i. Add I to the outgoing interface list of the (*,G) forwarding entry and set the timer for that interface (if there is no (*,G) entry, the router initializes one first),

      ii. For each (Si,G) entry associated with group G, if Si is not included in the prune list, and if I is not the iif then interface I is added to the *oif* list and the timers are reset for that interface in each affected entry,

      iii. If the (Si,G) entry is an RP-bit entry and its *oif* list is the same as (*,G) *oif* list, then the (Si,G,RPbit) entry is deleted,

      iv. The incoming interface is set to the interface used to send unicast packets to the RP in the (*,G) forwarding entry, i.e., RPF interface to the RP.

      v. The RP-list associated with the (*,G) entry is populated with the addresses found in the RP-address fields in the Join/Prune message.

(b) For each address Si in the join list whose RP-bit and WC-bit are *not* set, and for which there is no existing (Si,G) forwarding entry, the router initiates one. [8]

   i. The outgoing interface for (Si,G) is set to I (and I is added to the oif list for (*,G), if it exists). The incoming interface for (Si,G) is set to the interface used to send unicast packets to Si (i.e., the RPF neighbor).

   ii. If the interface used to reach Si is the same as the outgoing interface being built, I, this represents an error and the Join/Prune should not be processed.

   iii. If the source address in the join list of the Join/Prune message has its RP Annotated-bit set, the corresponding (S,G) state entry stores the address found in the RP-Address-1 field for G in the RP-Annotated field. The A-bit for interface I is set accordingly.

(c) For any Si included in the join list of the PIM-Join/Prune message, for which there is an existing (Si,G) forwarding entry,

   i. If the RP-bit is not set for Si listed in the Join/Prune message, but the RP-bit is set on the existing (Si,G) entry, the router clears the RP-bit on (Si,G) entry, sets the incoming interface to point towards Si for that (Si,G) entry, and sends a Join/Prune to the new incoming interface; and

   ii. The router adds I to the list of outgoing interfaces if I is not the same as the existing incoming interface; the timer for I is initialized.

   iii. The (Si,G) SPT bit is initialized to be cleared until data comes down the shortest path tree.

   iv. If the RP Annotated-bit (the A-bit) in Si's source-address flags-field in the join list is set, the address found in the RP-Address-1 field is copied into the RP-Annotated field in the (Si,G) state entry. The A-bit is set for the oif on which the Join/Prune message arrived. If the router is a DR, it also resets its Ack-Request timer for that group to suppress Ack-requests.

   v. If the RP-annotate bit (the A-bit) is cleared then the A-bit is cleared in the oif on which the Join/Prune message arrived. Also the RP-Annotated field is updated accordingly.

(d) For each address Si in the prune list,

   i. If there is an existing (Si,G) forwarding entry, the router schedules a deletion of I from the list of outgoing interfaces. If I is a multi-access LAN, the deletion is not executed until a timer expires; allowing for other downstream routers on the LAN to override the prune.

   ii. If the router has a current (*,G) forwarding entry, and if a (Si,G) RP-bit entry also exists then the (Si,G) RP-bit entry is maintained even if its outgoing interface list is null.

(e) For any Si in the prune list that has the RP-bit set:

   i. If (*,G) state exists, but there is no (Si,G) entry, an (Si,G,RP-bit) entry is created . The outgoing interface list is copied from the (*,G) entry, with the interface, I, on which the prune was received deleted. Packets from the pruned source, Si, match on this state and are not forwarded toward the pruned receivers.

   ii. If there exists a (Si,G) entry, with or without the RPbit set, the iif on which the prune was received, I, is deleted from the *oif* list, and the entry timer is reset.

[8] The router creates a (S,G) entry and copies all outgoing interfaces, excluding iif(S,G), from the (*,G) entry, if it exists. If a router does not copy all outgoing interfaces from the (*,G) entry, all receivers on RP-tree, downstream from outgoing interfaces other than the one newly added to (S,G), will not receive packets from source S. Data packets of S arriving from the RP will match the (S,G) entry instead of (*,G) entry, and will be dropped because the incoming interface is incorrect.

(f) When a DR receives a Join/Prune message for an (S,G) entry for a directly connected source, and the source's RP Annotated-bit is set to one, and the message contains a valid RP address in the group's RP-Address-1 field, the DR sets its RP-Register timer; this suppresses the sending of registers for that source. If the RP-Register timer expires, the A-bit is reset, and this causes subsequent packets from the source to be encapsulated and sent in Register messages to the active RP. If the RP-timer expires subsequent packets will trigger sending of Registers to the next RP in the RPlist.

2. If the received Join/Prune does not indicate the router as its target, then if the Join/Prune is for a (S,G) pair for which the router has an active (S,G) entry, and if the Join/Prune arrived an the *iif* for that entry. The router compares the IP address of the generator of the Join/Prune, to its own IP address.

   (a) If its own IP address is higher, the Joiner-bit in the (S,G) entry is set.
   (b) If its own IP address is lower, the Joiner-bit in the (S,G) entry is cleared, and the Joiner-bit timer is activated.

   After the timer expires the Joiner-bit is set indicating further periodic Join/Prunes should be sent for this entry. The Joiner-bit timer is reset each time a Join/Prune message is received from a higher-IP-addressed PIM neighbor.

   For any new (S,G) or (*,G) entry created by an incoming Join/Prune message, the Joiner-bit is set and the SPT-bit is cleared.

3.4 RP-Reachability

RP-Reachability messages are sent by the RP and processed by last hop routers on the shared RP-distribution tree. A router acts as an RP when it has (*,G) state with a null incoming interface and itself as the associated RP; this state is set up in response to (*,G) Join messages that indicate the router as the associated RP.

3.4.1 Sending RP-Reachability messages

The router sends the periodic RP-Reachability messages out all outgoing interfaces in the (*,G) entry. The default interval for this message is 90 seconds. The messages are addressed to the All-Routers Group (224.0.0.2) class D address and the message content includes the RP and G.

When an alternate active RP detects that a preferred RP is now reachable, it includes the address of the reachable, preferred RP in its RP-reachability messages. This is referred to as the active-RP-address field.

3.4.2 Receiving RP-Reachability messages

When a router receives an RP-Reachability message it does the following:

1. If the arriving interface for the RP-reachability message is not the same as the incoming interface in the (*,G) entry, drop the RP-Reachability message.

2. If the router is a last-hop router (i.e., has directly connected members), check if the active-RP-address included inside the PIM message is the same as the current RP being used. If it is the same, simply reset the corresponding RP-timer. If it is different, reset the RP-timer and update the (*,G) entry incoming interface to point to the now-reachable preferred RP indicated in the RP-Reachability message and trigger a (*,G) join toward that RP.

3.5 Register and Register-Ack

When a source first starts sending to a group its packets are encapsulated in PIM-Register messages and sent to the active RP. The RP sends Register-Ack messages towards the source(s); or if their data rate warrants source-specific paths, the RP sets up source specific state and starts sending (S,G) Join/Prune messages toward the source, with an annotation indicating that the Joins were initiated by the RP and act as an implicit Register-Ack.

3.5.1 Sending Registers and Receiving Register-Acks

Register messages are sent as follows:

1. When a DR receives a packet from a directly connected source, S:

   (a) If there is no corresponding (S,G) entry, the DR creates one with the Register-flag set to 1 and the RP address set according to RP-Group-mapping state in the DR. The Register-flag-timer is initialized to zero; the Register-flag-timer is non-zero only when the Register flag is set to 0. If there is no existing (*,G) or (Si,G) state for this group, the RP-timer and Ack-Request timers are initialized and the Ack-Request flag is set to 1.

   (b) If there is a (S,G) entry in existence, the DR simply resets the corresponding S-timer (entry timer).

   The Register-flag-timer is initialized to one-third the RP-timer when the Register flag for the (S,G) entry is set to zero (cleared). The Register-flag-timer and the Register-flag are reset by no-data Register-Acks for the particular source. They are also reset by Join/Prune messages with the RP-annotated bit set and RP-field indicating the current RP. IF and when the Register-flag-timer expires, the Register flag for that entry is set to one to reinstigate Register messages for that source.

2. If the new or previously-existing (S,G) entry has the Register-bit set, the data packet is encapsulated in a Register message and unicast to the active RP for that group. The data packet is also forwarded according to (S,G) state in the DR if the oif list is not null; since a receiver may join the SP-tree while the DR is still registering to the RP.

   (a) If the RP-Group-mapping state's Ack-request flag is set, the DR sets the Ack-request flag in the Register message and clears the Ack-request flag in the RP-Group-mapping state. It also resets the Ack-request timer.

3. If the (S,G) entry has the Register-flag cleared, the data packet is not sent in a Register message, it is just forwarded according to the (S,G) oif list.

4. If the DR's Ack-Request timer expires for some group, G, the following actions are taken if the RP-Group-mapping timer has not expired:

   (a) The DR schedules sending of a null-Register message (i.e., a Register message with no encapsulated data and the Ack-Request and no-data flags set to 1.) The null-Register message is scheduled for sending after some short delay. The DR sets the Ack-Request flag for the group. This delay is introduced to take advantage of piggybacking the Ack-Request in a pending Register with encapsulated data.

   (b) If the DR flag for the group is not cleared within the short time period scheduled, the DR sends the null-Register message.

In this way, the Ack-request timer is used to drive periodic probing of the RP using the Ack-request flag in either data-driven Registers or null-Registers. The Ack-Request timer is reset, and null-Register messages are suppressed, by indications of RP reachability (Register-Acks or Join/Prune messages with the RP-annotated bit set by the RP). The Ack-request timer is initialized to one third of the RP-timer initialization value. The RP-Group-mapping timer indicates that the DR has directly connected sources interested in sending to the group. The RP-Group-mapping timer is reset by IGMP RP-Report messages sent by directly-connected hosts, with the source-flag set in the RP-Report message.

5. If the DR's RP-timer expires for some group, G, the following actions are taken:
   (a) The DR assumes that the current RP is not reachable and chooses the next RP on the RP-list. Subsequent Register messages are sent to the newly selected RP. The RP-timer is reset. In addition, the Register flags and Register-timers for all existing (Si,G) entries are reinitialized. The RP-list is treated as a ring so that the first RP is tried again, following the last RP on the list.
   (b) The subsequent data packets or null-Register messages are sent to the new RP.

   The RP-timer is reset by indications of RP reachability (Register-Acks or Join/Prune messages with the RP-annotated bit set by that RP.)

The DR processes Register-Ack messages as follows:

1. If the RP-address is different from the RP address currently used by the DR, the DR sets the active RP for that group to that indicated in the Register-Ack and resets the Register-flag in corresponding (S,G) entries. If the indicated RP-address is not a valid IP unicast address it should be ignored.
2. The DR resets the Ack-Request-timer and RP-timer for the corresponding group.
3. If the no-data flag is set, the DR clears the Register-flag and initializes the Register-flag-timer in the corresponding (S,G) entry(ies).

When a Register-flag-timer expires, the corresponding entry(ies) Register flag is set to 1 to reinstigate encapsulation of data packets in Register messages.

3.5.2 Receiving Register Messages and Sending Register-Acks

When a router (i.e., the RP) receives a Register message, the router does the following:

1. Decapsulates the data packet, and checks for a corresponding (S,G) entry.
   (a) If a (S,G) entry exists and the packet arrived from the decapsulation process, the packet is forwarded but the SPT bit is left cleared (0). If the SPT bit is 1, the packet is dropped.
   (b) If there is no (S,G) entry, but there is a (*,G) entry, the packet is forwarded according to the (*,G) entry.
   (c) If there is no G related entry, the RP initializes one with a null oif list and the iif null. A timer for the entry is also initialized.

   The (S,G) state timer is reset by packets arriving from that source to that group. If the Register message contains a null-data portion, the (S,G) state timer is still reset.

2. If the Ack-Request flag is set in the Register message, or if the matching (S,G) or (*,G) state contains a null oif list, the RP unicasts a Register-Ack message to the source of the Register message. If the relevant entry, either (S,G) or (*,G), has a null oif list, then the no-data flag is set; in the latter case, the source-address field is set to the wildcard value (all 0's). This message is not processed by intermediate routers, hence no (S,G) state is constructed between the active RP and the source. Register-Acks are rate limited.

3. If the Register message arrival rate warrants it and there is no existing (S,G) entry, the RP sets up a (S,G) forwarding entry with the outgoing interface list, excluding iif(S,G), copied from the (*,G) outgoing interface list, its SPT-bit is initialized to 0. The (S,G) state entry includes the RP-Address in the RP-Annotated field. A timer is set for the (S,G) entry and this timer is reset by receipt of data packets for (S,G). The (S,G) entry causes the RP to send a Join/Prune message for the indicated group towards the source of the register message. The Join/Prune message includes the RP's address in the RP-Address-1 field for that group. The Join/Prune message includes the source's address in the Join list with its RP Annotated-bit set to 1.

   If the (S,G) oif list becomes null, Join/Prune messages will not be sent towards the source, S.

4. If the currently active RP is not the preferred RP, it periodically polls the preferred RP(s) (all the RPs that appear before it in the ordered RP-list). When one of the preferred RPs becomes reachable, the active alternate RP unicasts a Register-Ack to the sources' first-hop routers; the message contains the address of the now-reachable and preferred RP in the active-RP-address field. See the section on RP Polling for more details.

3.6 Poll and Poll-Response

The Poll message is used by an alternate RP to check the status of preferred RPs. The Poll-Response message is sent from a recovered (or now reachable) preferred RP to the currently-active alternate-RP to notify it of this recovery.

3.6.1 Sending Poll

The following events trigger sending Poll messages:

1. When a PIM router receives a Join/Prune message with its address in the join list with the RP-bit and WC-bit set (hence it knows it is the active RP), the router starts sending periodic Poll messages to preferred RPs; i.e. the RPs that are before it in the ordered RP-list included in the Join/Prune message (note that the first RP on the list does not send Polls).

2. When a PIM router receives a Register message, it starts sending periodic Poll messages to the preferred RPs.

Poll messages are sent with the Poll bit set.

3.6.2 Receiving Poll and Sending Poll-Response

When a PIM router receives a Poll message, it clears the Poll bit in the message (hence, the message becomes a Poll-Response) and sends the message back to its source.

3.6.3 Receiving Poll-Response

When the active-alternate RP receives a Poll-Response message, it performs the following:

1. Includes the RP-Address of the now-reachable and preferred RP in the RP-Reachability messages sent down the (*,G) tree to receivers.
2. Includes the RP-Address of the now-reachable and preferred RP in Register-Ack messages unicast to sources.

A Register-Ack message is triggered when the active RP finds that a preferred-RP is reachable.

3.7 Multicast Data Packet Forwarding

Processing a multicast data packet involves the following steps:

1. Lookup forwarding state based on a longest match of the source address, and an exact match of the destination address in the data packet and compare the RPF check on the source address in the packet header with the *iif* specified in the forwarding entry.
2. If the packet arrived on the interface found in the matching-entry's *iif* field:
   (a) Forward the packet to the *oif* list for that entry and reset the entry's timer.
   (b) If the entry's SPT-bit is cleared, set the SPT-bit for that entry. If (*,G) also exists and their incoming interfaces are different, trigger a (S,G) prune with RP-bit set towards the active RP.
   (c) If the source of the packet is a directly-connected host and the router is the DR on a multi-access network, check the Register flag associated with the (S,G) entry. If it is set, then the router encapsulates the data packet in a register message and sends it to the active RP.

   This covers the common case of a packet arriving on the RPF interface to the source or RP and being forwarded to all joined branches. It also detects when packets arrive on the SP-tree, and triggers their pruning from the RP-tree. If it is the DR for the source, it sends data packets encapsulated in PIM-Registers to the RPs.
3. If the packet matches to an entry but did not arrive on the the interface found in the entry's *iif* field, check the SPT-bit of the entry. If the SPT-bit is set, drop the packet. If the SPT-bit is cleared, then lookup the (*,G) entry for the packet. If the packet arrived on the *iif* found in (*,G), forward the packet to the *oif* list of the (S,G) entry. This covers the case when a data packet matches on a (S,G) entry for which the SP-tree has not yet been completely established upstream.
4. If the packet does not match to any entry, but the source of the data packet is a local, directly-connected host, and if the router is the DR on a multi-access LAN and knows of the active RP associated with the destination group, G, then the DR checks the register flag associated with the local sender (if there is no such a register flag, a new register flag, associated with the local sender, is created and set), the data packet is encapsulated in a register message and sent to the active RP.
5. If the packet does not match to any entry, and it is not a local host or the router is not the DR, drop the packet.

3.8 Data triggered switch to shortest path tree (SP-tree)

When a (*,G) entry is created, a data rate counter may be initiated at the last-hop routers. The counter is incremented with every data packet received for directly connected members of an SM group, if the longest match is (*,G). If and when the data rate for the group exceeds a certain configured threshold (t1), the router initiates 'source-specific' data rate counters for the following data packets. Then, each counter for a source, is incremented when packets matching on (*,G) are received from that source. If the data rate from the particular source exceeds a configured threshold (t2), a (S,G) entry is created and a Join/Prune message is sent towards the source. If the RPF interface for (S,G) is *not* the same as that for (*,G), then the SPT-bit is cleared in the (S,G) entry. Other configured rules may be enforced to cause or prevent establishment of (S,G) state.

3.9 Assert

Asserts are used to resolve which of the parallel routers connected to a multi-access LAN is responsible for forwarding packets onto the LAN.

3.9.1 Sending Asserts

The following Assert rules are provided when a multicast packet is received on an outgoing multi-access interface of an existing (S,G) entry:

1. Do unicast routing table lookup on source IP address from data packet, and send assert on interface for source IP address in data packet; include metric preference of routing protocol and metric from routing table lookup.
2. If route is not found, use metric preference of 0x7fffffff and metric 0xffffffff.
3. When an assert is sent for a (*,G) entry, the first bit in the metric preference (the RP-bit) is set to 1, indicating the data packet is routed down the RP-tree.

Asserts are rate-limited by the router.

3.9.2 Receiving Asserts

When an assert is received the router performs a longest match on the source and group address in the assert message. The router checks the first bit of the metric preference (RP-bit). If the RP-bit is set, the router does a match on (*,G) entries, otherwise, the router matches (S,G) entries. If the interface that received the Assert message is in the *oif* list of the matched entry, then this assert is targeted for this router and the message is processed as follows:

1. Compare the metric received in the Assert with the one the router would have advertised in an assert. Note that, the metric preference should be treated as the high-order part of an assert metric comparison. If the value in the assert is less than the router's value, delete the interface from the entry. If the value is the same, compare IP addresses, if the routers address is less than the assert sender, delete the interface.
2. If the router has won the election and there are directly connected members on the multi-access LAN, the router keeps the interface in its outgoing interface list. It acts as the forwarder for the LAN.

3. If the router won the election but there are no directly connected members on the multi-access LAN, the router schedules to delete the interface. The LAN might be a stub LAN with no members (and no downstream routers). If no subsequent Join/Prunes are received, the router deletes the interface from the outgoing interface list; otherwise it keeps the interface in its outgoing interface and acts as the forwarder for the LAN.

The winning router should send out an assert message including its own metric to that outgoing interface, so the other router will prune that interface from its forwarding entry. Also, when an assert is received, the router performs an exact match based on the source address, group address and the RP-bit of the metric preference in the assert message. Note that, this is not a longest match, only exact state will be matched. If there is no such state, then the router drops the assert message. Otherwise, If the interface that received the assert matches the incoming interface of the exactly matched entry, then the assert message is processed as follows:

1. Downstream routers will select the upstream router with the smallest metric as their RPF neighbor. If two metrics are the same, the highest IP address is chosen to break the tie.

2. If the downstream routers have downstream members, they must schedule a join to inform the upstream router that packets should be forwarded on the multi-access network. This will cause the upstream forwarder to cancel its delayed deletion of the interface.

3.10 Candidate-RP-Advertisements

Candidate-RP-Advertisements are low frequency PIM-messages sent by PIM routers willing to become RPs. The messages are sent to a well-known multicast group. Group initiators use these advertisements to build the RP-list for the group.

Candidate-RP-Advertisements carry group address and group mask fields. This enables the advertising router to limit the advertisement to a certain range or scope of groups. The router may enforce this scope acceptance when receiving Registers or Join/Prune messages.

3.11 Processing Timer Events

Editors Note: This subsection needs some more work to be complete. We decided we should have a separate section on timer processing but we have a bit more work to do before this section is complete, ie. before ALL timers used in PIM are described here in detail. Timers are also discussed individually in the sections that pertain to the protocol messages that they trigger/affect.

In this subsection we mention some critical timer events that are not always associated with the receipt or sending of messages and therefore are not fully covered by earlier subsections.

Each (S,G) and (*,G) entry has timers associated with it. There are multiple timers maintained: one for the multicast routing entry itself and one for each interface in the outgoing interface list.

Timers on entries are handled as follows:

1. The **S-timer** of a (S,G) entry is reset whenever data packets for (S,G) are forwarded, or when a PIM-Register is received from S.

2. The entry-timer for a (*,G) entry is reset when any of its oif timers gets reset.

3. The S-timer for a (S,G) RP-bit entry is reset whenever a (S,G) prune with RP-bit set is received.

Each timer expires after 3 times the refresh period; a typical value is 3 minutes (because a typical Join/Prune refresh interval is 1 minute.)

The RP-Group-mapping timer is a group-specific, not source-specific, timer, that is initialized when a DR first receives an RP-Report message for a particular group. The timer is reset when the DR receives an IGMP RP-Report for that group. So long as this timer is non–zero, the DR will enable periodic null-Register messages to keep track of RP liveness.

A timer is also maintained for each outgoing interface listed in each (S,G) or (*,G) entry. Each oif timer is managed as follows.

1. The timer is set when the interface is added.
2. The timer is reset each time a Join/Prune message or an IGMP membership report for G is received on that interface for that forwarding entry (i.e., (*,G)).
3. When a timer is reset for an outgoing interface listed in a (*,G) entry, the timers are reset for that interface, in all existing (S,G) entries whose oif list contains that interface. Because some of the outgoing interfaces in (S,G) entry are copied from the (*,G) outgoing interface list, they may not have explicit (S,G) join messages from some of the downstream routers (i.e., where members are joining to the (*,G) tree only). [9]
4. When an outgoing interface timer expires, the corresponding outgoing interface is deleted from the outgoing interface list.

A deletion timer is used to schedule deletion of multicast forwarding entries. Entries may be scheduled for deletion when their oif lists become null:

1. When the oif list of an (S,G) entry becomes null, and the RP-bit is not set to 1, the entry is scheduled for deletion.

   [10] Once the (S,G) is timed out, it may be recreated when the next Join/Prune arrives.

   When the oif list for a (*,G) entry is null in a router that is not a DR or the RP, the entry is deleted.
2. When the oif list for a (*,G) entry is null in the router that is the RP for that entry, the entry is scheduled for deletion (to allow time for polling preferred RPs).
3. When the oif list for a (*,G) entry is null in a router that is the DR, and the RPF neighbor to the RP is the LAN, then the (*,G) entry is kept alive even though the oif list is null.
4. When a (*,G) entry is deleted, all associated (S,G,RPbit) entries are also deleted.

The RP-timer is a timer associated with the active RP per group.

1. When an RP-Reachability message is received at the receivers' last hop routers the RP timer is reset. RP-Reachability messages contain the time-out period. The RP timer must be set to this value.

---

[9] If there are sources in the prune list of the (*,G) join, then the timers for arriving interface will first be reset for those sources, and then this interface will be deleted from these same entries; producing a correct result, even though the updating of the timers was unnecessary. An implementation could optimize this by checking the prune list before processing the join list.

[10] (S,G) entries with the RP-bit set, i.e., (S,G) RP-bit entries, are kept alive by receipt of Prunes. We do not want to delete such entries if a (*,G) entry exists; otherwise, data packets will travel down both the RP-tree and SP-tree. While this would not result in periodic duplicates (because of the RPF check), it would waste network bandwidth.

2. When a Register-Ack or Join/Prune with the RP-annotate bit and RP address included is received at a DR with the corresponding (S,G) state, the associated RP timer is updated; the RP timer is set to 270 seconds, or to the Holdtime in the Join/Prune message, respectively.

Editors Note: The Assert timer sections were added recently. Will be sanity-checked for next I-D submission.

Routers on multiaccess LANs have Assert Timers. This timer is set in the downstream router(s) when one of the upstream routers on the LAN wins an Assert and becomes the upstream neighbor, in conflict with the unicast routing table's RPF upstream neighbor. When this timer expires, the downstream router should change its RPF neighbor back to the unicast routing table's RPF neighbor so as to reflect topology changes. *TBD*

Another timer associated with Asserts is the Assert Rate-limit timer referred to in the section on Processing Assert messages. The Assert Rate-limit timer is reset whenever an Assert message is sent. An Assert message is not sent for a particular oif unless the Assert Rate-limit timer expires.

4 Packet Formats

RFC-1112, see [5], specifies two types of IGMP packets for hosts and routers to convey multicast group membership and reachability information. An IGMP Host-Query packet is transmitted periodically by routers to ask hosts to report multicast groups of which they are members. An IGMP Host- Report packet is transmitted by hosts in response to received queries advertising group membership.

This section introduces new types of IGMP packets that are used by PIM routers. All PIM control messages are encoded in IGMP messages.

4.1 IGMP Fixed Header

The fixed header packet format is:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |      Code     |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Address                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version This memo specifies version 1 of IGMP.

Type There are nine types of IGMP messages:

1 = Host Membership Query
2 = Host Membership Report
3 = Router DVMRP Messages
4 = Router PIM Messages
5 = Cisco Trace Messages
6 = New Host Membership Report
7 = Host Membership Leave
14= Mtrace Response
15= Mtrace Request

Code Codes for specific message types. Used only by DVMRP and PIM. PIM codes are:

0 = Router-Query
1 = Register
2 = Register-Ack
3 = Join/Prune
4 = RP-Reachability
5 = Assert
6 = Graft
7 = Graft-Ack
8 = Candidate-RP-Advertisement
9 = Poll

Checksum The checksum is the 16-bit one's complement of the one's complement sum of the entire IGMP message. For computing the checksum, the checksum field is zeroed.

Address PIM Version field when IGMP type is PIM.

4.2 PIM Fixed Header

The PIM fixed header carries the PIM version number, in addition to a reserved field and address length specifier fields.

```
 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|               Reserved                | Addr length   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

PIM Ver PIM Version number is 2.

Reserved Transmitted as zero, ignored on receipt.

Addr length Address length in bytes. Throughout this section this would indicate the number of bytes in the Address field of an address.

4.3 Encoded Source and Group Address formats

1. Unicast address: Only the address is included. The length of the unicast address in bytes is specified in the 'Addr length' field in the header.

2. Encoded-Group-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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Reserved  |  Mask Len     | Group multicast Address ...   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| ...Group multicast Address ...|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+~+~+~+~+~+~+
```

Reserved Transmitted as zero. Ignored upon receipt.

Mask Len The Mask length is 8 bits. The value is the number of contiguous bits left justified used as a mask which describes the address. It is less than or equal to Addr length * 8. If the message is sent for a single group then the Mask length should equal Addr length * 8. In version 2 of PIM, it is strongly recommend that this field be set to 32 for IPv4 and 128 for IPv6.

Group multicast Address contains the group address, and has number of bytes equal to that specified in the Addr length field.

3. Encoded-Source-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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Rsrvd |A|S|W|R|  Mask Len     | Source Address ...            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  ...  Source Address          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+~+~+-+
```

Reserved Transmitted as zero, ignored on receipt.

A,S,W,R See section 4.7 for details.

Mask Length Mask length is 8 bits. The value is the number of contiguous bits left justified used as a mask which describes the address. The mask length must be less than or equal to Addr Length * 8. If the message is sent for a single source then the Mask length should equal Addr length * 8. In version 2 of PIM, it is strongly recommend that this field be set to 32 for IPv4.

Source Address The address length is indicated from the Addr length field at the beginning of the header. For IPv4, the address length is 4 octets.

4.4 PIM-Query Message

It is sent periodically by PIM routers on all interfaces.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |      Code     |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver|                Reserved               | Addr length   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Reserved          |           Holdtime            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version, Type, Code, Checksum, PIM Version Described above.

Reserved Transmitted as zero, ignored on receipt.

Addr length not used.

Holdtime The amount of time a receiver should keep the neighbor reachable, in seconds.

4.5 PIM-Register Message

It is sent by the Designated Router (DR) to the active RP when a multicast packet needs to be transmitted on the RP-tree. Source IP address is set to the address of the DR, destination IP address is to the RP's address.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |      Code     |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver|                Reserved               | Addr length   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|D|Q|             Reserved                              |RP-Cnt |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Unicast-RP-Address-1                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               . .                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Unicast-RP-Address-m                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                       Multicast data packet                   ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version, Type, Code, Checksum, PIM Version Described above.

Addr length The unicast address length in bytes. Specifies the address length of the Unicast-RP-Address fields.

D The data flag, when cleared indicates no-data included in the Multicast data packet section. The D flag is cleared in null-Registers.

Q The Ack-Request flag, is a 1 bit value. When set, signals Register-Acks to be sent in response. The Q flag is set periodically to trigger periodical Register-Acks in response.

RP-Cnt The number of RP-Addresses include in the RPlist.

Unicast-RP-Address-1 .. m The ordered RPlist, listed in order of preference.

Multicast data packet The original packet sent by the source. For periodic sending of registers with the D flag cleared, this part contains only the IP header.

4.6 PIM-Register-Ack Message

It is triggered by the Ack-Request flag set in a Register message. A Register-Ack is unicast from the active RP to the sender of the Register message. Source IP address is the address to which the register was addressed. Destination IP address is the source address of the register message.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |     Code      |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver|              Reserved                 | Addr length   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Encoded-Group Address                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Unicast-Source Address                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Unicast-Active-RP-Address                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|N|                          Reserved                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version, Type, Code, Checksum, PIM Version, Addr length Described above.

Encoded-Group Address Format described above. Note that for Register-Acks the Mask len field should contain Addr length * 8 (32 for IPv4), if the message is sent to a single group.

Unicast-Source Address IP host address of source from multicast data packet in register. The length of this field in bytes is specified in the Addr length field.

Unicast-Active-RP-Address The address of the now reachable and preferred RP. The length of this field in bytes is specified in Addr length field. If included, and different than the source of the Register-Ack, then the sender's DR would know to register to the RP that is given in the RP-Address field. If this field does not contain a valid IP unicast address it should be ignored.

N No-Data flag. A bit, when set informs the source not to send any data in the Registers.

4.7 Join/Prune Message

It is sent by routers towards upstream sources and RPs. A join creates forwarding state and a prune destroys forwarding state. Joins are sent to build shared trees (RP trees) or source trees (SPT). Prunes are sent to prune source trees when members leave groups as well as sources that do not use the shared tree.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |     Code      |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver|                Reserved               | Addr length   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Unicast-Upstream Neighbor Address                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Reserved     | Num groups    |          Holdtime             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Encoded-Multicast Group Address-1                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Reserved   |RP-Cnt |Number of Join Srcs|NumberOf Prune Srcs|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Unicast-RP Address-1                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             .                                 |
|                             .                                 |
|                             .                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Unicast-RP Address-m                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Encoded-Join Source Address-1                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              .                                |
|                              .                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Encoded-Join Source Address-n                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Encoded-Prune Source Address-1                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              .                                |
|                              .                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Encoded-Prune Source Address-n                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            .                                  |
|                            .                                  |
|                            .                                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Encoded-Multicast Group Address-n              |
```

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Reserved      |RP-Cnt |Number of Join Srcs|NumberOf Prune Srcs|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Unicast-RP Address-1                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            .                                  |
|                            .                                  |
|                            .                                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Unicast-RP Address-m                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Encoded-Join Source Address-1                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             .                                 |
|                             .                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Encoded-Join Source Address-n                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Encoded-Prune Source Address-1                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             .                                 |
|                             .                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Encoded-Prune Source Address-n                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version, Type, Code, Checksum, PIM Version Described above.

Addr Length The length in bytes of the encoded source addresses in the join and prune lists and the unicast RP-Addresses.

Upstream Neighbor Address The IP address of the RPF or upstream neighbor.

Reserved Transmitted as zero, ignored on receipt.

Holdtime The amount of time a receiver should keep the Join/Prune state alive, in seconds.

Number of Groups The number of multicast group sets contained in the message.

Encoded-Multicast group address For format description see section 4.3.

RP-Cnt The RP count field contains the number of RP addresses included in the message for a specific group. For (*,G) Join/Prune messages (RP count >=1); depending on the number of RP's in the RPlist. For (S,G) Join/Prune messages sent from the RP to a source, the RP count is set to 1. For (S,G) Join/Prune messages in which all sources in the Join list have their RP Annotated bits (A-bits) set to 0, the RP-Cnt is set to 0.

Unicast-RP Address-1 .. m This is a list of the RPs. RPs are listed in preference order received.

Number of Join Sources Number of join source addresses listed for a given group.

Join Source Address-1 .. n This list contains the sources that the sending router will forward multicast datagrams for if received on the interface this message is sent on.

See format section 4.3. The fields explanation for the Encoded-Source-Address format follows:

Reserved Described above.

A RP Annotated-bit. When set, the RP Address is annotated in corresponding (S,G) entry. The A bit is always set to 0 for sources in the prune list.

S The Sparse bit is a 1 bit value, it is used by routers on the shortest path tree to indicate the group is in sparse-mode (since they do not know about any RPs for the group). This indicates to receivers to send periodic Join/Prune messages towards the source. When set to 1, the (S,G) should be treated in sparse-mode, otherwise, it should be treated in dense-mode.

W The WC bit is a 1 bit value. If 1, the join or prune applies to the (*,G) entry. If 0, the join or prune applies to the (S,G) entry where S is Source Address. Joins and prunes sent towards the RP should have this bit set.

R The RP bit is a 1 bit value. If 1, the information about (S,G) is sent towards the RP. If 0, the information should be sent about (S,G) toward S, where S is Source Address.

Mask Length, Source Address Described above.

Represented in the form of $< WCbit >< RPbit >< Mask\ length >< Source\ address >$:

A source address could be a host IP address :

$$< 0 >< 0 >< 32 >< 192.1.1.17 >$$

A source address could be the RP's IP address :

$$< 1 >< 1 >< 32 >< 131.108.13.111 >$$

A source address could be a subnet address to prune from the RP-tree :

$$< 0 >< 1 >< 28 >< 192.1.1.16 >$$

A source address could be a general aggregate :

$$< 0 >< 0 >< 16 >< 192.1.0.0 >$$

Number of Prune Sources Number of prune source addresses listed for a group.

Prune Source Address-1 .. n This list contains the sources that the sending router does not want to forward multicast datagrams for when received on the interface this message is sent on. See format below.

4.8 PIM-RP-Reachability Message

Each RP will send RP-Reachability messages to all routers on its distribution tree for a particular group. These messages are sent so routers can detect that an RP is reachable. Routers that have attached host members for a group will process the message.

The RPs will address the RP-Reachability messages to 224.0.0.2 (All-Routers-Group). Routers that have state for the group with respect to the RP distribution tree will propagate the message. Otherwise, the message is discarded.If an RP address timer expires, the router should attempt to send a PIM join message towards an alternate RP provided for that group if one is available.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |     Code      |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver|               Reserved                | Addr length   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Encoded-Group Address                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Unicast-RP Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Reserved          |       Holdtime                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version, Type, Code, Checksum, PIM Version, Addr length Described above.

Encoded-Group Address The group address the RP is associated with. Format described earlier.

Unicast-RP Address The rendezvous point IP address of the sender. If the RP Address field is different than the currently active RP, then the member's DR should join to the RP given in that field. If this field does not contain a valid IP unicast address it should be ignored. The length of this field in bytes is specified in Addr length.

Reserved Transmitted as zero, ignored on receipt.

Holdtime The amount of time in seconds receivers of this message should consider the RP reachable.

4.9 PIM-Assert Message

The PIM-Assert message is sent when a multicast data packet is received on an outgoing interface corresponding to the (S,G) or (*,G) associated with the source.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |     Code      |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver|               Reserved                | Addr length   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Encoded-Group Address                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Unicast-Source Address                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|R|                         Metric Preference                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Metric                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version, Type, Code, Checksum, PIM Version, Addr length Described above.

Encoded-Group Address The group address to which the data packet was addressed, and which triggered the Assert. Format previously described.

Unicast-Source Address Source IP address from IP multicast datagram that triggered the Assert packet to be sent. The length of this field in bytes is specified in Addr length.

R RP bit is a 1 bit value. If the IP multicast datagram that triggered the Assert packet is routed down the RP tree, then the RP bit is 1; if the IP multicast datagram is routed down the SPT, it is 0.

Metric Preference Preference value assigned to the unicast routing protocol that provided the route to Host address.

Metric The unicast routing table metric. The metric is in units applicable to the unicast routing protocol used.

4.10 PIM-Graft Message

Used in dense-mode. Refer to PIM dense mode specification.

4.11 PIM-Graft-Ack Message

Used in dense-mode. Refer to PIM dense mode specification.

4.12 Candidate-RP-Advertisement

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |     Code      |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver|               Reserved                | Addr length   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Intended-Hop-Count            |             Holdtime          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Encoded-Group Address                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version, Type, Code, Checksum, PIM Version Described above.

Addr length not used in this message.

Intended-Hop-Count This field is copied from the original TTL field when the advertisement is originated. It is not modified by intermediate routers.

Holdtime The amount of time the advertisement is valid. This field allows advertisements to be aged out.

Encoded-Group Address The group address the RP is associated with. Format previously described.

4.13 PIM-Poll and Poll-Response Message

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |     Code      |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver|               Reserved                | Addr length   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|P|                               Reserved                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version, Type, Code, Checksum, PIM Version Described above.

Addr length not used in this message. Transmitted as zero, and ignored upon receipt.

P The poll bit. When set indicates a Poll message, and when cleared indicates a Poll-Response.

5 Pseudocode

Editors Note: This section is still in progress. In the future the pseudocode will be available by anonymous ftp at catarina.usc.edu:pub/estrin/pim/pim.sm.pseudocode.

6 Acknowledgments

Tony Ballardie, Scott Brim, Jon Crowcroft, Bill Fenner, Paul Francis, Joel Halpern, Horst Hodel, Stephen Ostrowski, Dave Thaler, and Lixia Zhang provided detailed comments on previous drafts. The authors of CBT[8] and membership of the IDMR WG provided many of the motivating ideas for this work and useful feedback on design details.

This work was supported by the National Science Foundation, ARPA, cisco Systems and Sun Microsystems.

References

[1] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, L. Wei, P. Sharma, and A. Helmy. Protocol independent multicast (pim) : Motivation and architecture. *Internet Draft*, May 1995.

[2] S. Deering, D. Estrin, D. Farinacci, B. Fenner, V. Jacobson, and A. Helmy. Interoperability architecture and mechanisms for pim-sm. *Internet Draft*, June 1995.

[3] S. Deering, D. Estrin, D. Farinacci, and V. Jacobson. Protocol independent multicast (pim), dense mode protocol : Specification. *Internet Draft*, March 1994.

[4] D. Waitzman S. Deering, C. Partridge. Distance vector multicast routing protocol, nov 1988. RFC1075.

[5] S. Deering. Host extensions for ip multicasting, aug 1989. RFC1112.

[6] S. Deering. Igmp. *???*, November 1994.

[7] J. Moy. Multicast extension to ospf. *Internet Draft*, September 1992.

[8] A. J. Ballardie, P. F. Francis, and J. Crowcroft. Core based trees. In *Proceedings of the ACM SIGCOMM*, San Francisco, 1993.