Protocol Independent Multicast (PIM): Protocol Specification

Stephen Deering

Xerox PARC
3333 Coyoty 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

Computer Science Department
University of Southern California
Los Angeles, CA 90089
charley@catarina.usc.edu

Liming Wei

Computer Science Department
University of Southern California
Los Angeles, CA 90089
lwei@catarina.usc.edu

draft-ietf-idmr-pim-spec-01.ps

January 11, 1995

Status of This Memo

---

[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 (SPT) and contributed to protocol development as part of his simulation effort. Estrin, Liu, and Wei were supported by grants from the National Science Foundation and Sun Microsystems.

1 Introduction

This document describes protocols for efficiently routing to multicast groups that may span wide-area (and inter-domain) internets. We refer to the approach as Protocol Independent Multicast (PIM) because it is not dependent on any particular unicast routing protocol.

This document describes the protocol details. For the motivation behind the design and a description of the architecture, see [1].

Section 2 summarizes PIM operation in both Sparse Mode (SM) and Dense Mode (DM). It describes the protocol from the perspective of the overall network and how the participating routers interact to create and maintain the multicast distribution tree. Section 3 describes PIM 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 message type; for each message type we describe its contents, its generation, and its processing. Section 4 provides packet format details and section 5 provides pseudocode that corresponds to the functions described in section 3; however it is just for illustration.

Editors Note: the next version of the specification will include (1) a few small protocol changes to accommodate source-specific pruning off of the RP-tree, and (2) an updated and more detailed discussion of PIM/non-PIM interaction. Section 4 is authoritative.

2 PIM Protocol Overview

In this section we provide an overview of the architectural components of PIM. For clarity, we describe the general behavior of PIM-Dense Mode and PIM-Sparse Mode separately. However, the detailed protocol mechanisms developed to realize sparse and dense mode functionality are described in an integrated manner in subsequent sections. We also describe special mechanisms used by both PIM-DM and PIM-SM when operating over multi-access networks.

2.1 PIM-Dense Mode (PIM-DM)

PIM-DM forwards data packets onto all outgoing interfaces (except the expected incoming interface) until pruning occurs. Once truncation occurs, pruning state is maintained in routers that are not on the steady-state distribution tree, and packets are only forwarded onto outgoing interfaces (*oif*) that in fact reach downstream members.

The rest of this subsection describes the interaction between routers in creating dense mode multicast distribution tree state.

2.1.1 Leaf network detection

In DVMRP, poison reverse information tells a router that other routers on the shared LAN use the LAN as their incoming interface (*iif*). As a result, even if a router for that LAN does not hear any IGMP Host-Reports for a group, the router will know to continue to forward multicast data packets to that group, and *not* to send a prune message to its upstream neighbor.

Since PIM does not rely on any unicast routing protocol mechanisms, this problem is solved by using prune messages sent upstream on a LAN. If a downstream router on a LAN determines that it has no more downstream members for a group, then it can multicast a prune message on the LAN.

A last-hop router detects that there are no members downstream when it is the only active router on a network and there are no IGMP Host-Report messages received from hosts. It determines there are no other routers by not receiving PIM-Query messages.

If an (S,G) entry contains an empty outgoing interface list (i.e., an (S,G) negative cache entry), a prune is sent upstream. Prune information is flushed periodically. This (or a loss of state) causes the packets to be sent in reverse path forwarding (RPF) mode again which in turn triggers prune messages.

When a prune message is sent on an upstream LAN, it is data link multicast and IP addressed to the all routers group address 224.0.0.2. The router to process the prune will be indicated by inserting its address in the "Upstream Neighbor Address" field of the message. The address is obtained by an RPF lookup from the unicast routing table. When the prune message is sent, the expected upstream router will schedule a deletion request of the LAN from its outgoing interface list for the (S,G) entry in the prune list. The suggested delay time before deletion should be greater than 3 seconds. Prunes received on point-to-point links can prune right away without scheduling a deletion request.

Note the special case for equal-cost paths. When an upstream router is chosen by an RPF lookup there may be equal-cost paths to reach the source. The higher IP addressed system is always chosen. If the unicast routing protocol does not store all available equal-cost paths in the routing table, the "Upstream Neighbor Address" field may contain the address of the wrong upstream router. To avoid this situation, the "Upstream Neighbor Address" field may optionally be set to 0.0.0.0 which means that all upstream routers (the ones that have the LAN as an outgoing interface for the (S,G) entry) may process the packet.

Other routers on the LAN will hear the prune message and respond with a join if they still expect multicast datagrams from the expected upstream router. The PIM-Join message is data link multicast and IP addressed to the all routers group address 224.0.0.2. The router to process the join will be indicated by inserting its address in the "Upstream Neighbor Address" field of the message. The address is determined by an RPF lookup from the unicast routing table. When the expected router receives the join message, it will cancel the deletion request.

Routers will randomly generate a join message delay timer. If a join is heard from another router before a router sends its own, it will cancel sending its own join. This will reduce traffic on the LAN. The suggested join delay timer should be from 1 to 3 seconds.

If the expected upstream router does not receive any PIM-Join messages before the scheduled time for the deletion request expires, it deletes the outgoing LAN interface from the (S,G) multicast forwarding entry.

2.1.2 New members joining an existing group

If a router is directly connected to a host that wants to become a member of a group, the router may send a PIM-Graft message towards known sources. This reduces join latency indicated by the relatively large timeout value suggested for prune information.

If a receiving router has state for group G, it adds the interface on which the IGMP Host-Report or PIM-Graft was received for all known (S,G). If the (S,G) entry has an empty outgoing interface list, the router sends a PIM-Graft message upstream towards S.

If routers have no group state, they do nothing since dense-mode PIM will deliver a multicast datagram to all interfaces when creating state for a group.

If a router receives a PIM-Graft message on the incoming interface for the associated (S,G) entry, the router will not add the arriving interface to the outgoing interface list.

The PIM-Graft message uses a positive acknowledgment strategy. Senders of PIM-Graft messages unicast them to their upstream RPF neighbors. The neighbor processes each (S,G) and immediately acknowledges each (S,G) in a PIM-Graft-Ack message. This is relatively easy, since the receiver simply changes the IGMP code from PIM-Graft to PIM-Graft-Ack, recomputes the checksum, and unicasts the modified packet back to the source router. The sender periodically retransmits the PIM-Graft message for any (S,G) that has not been acknowledged. Note that the sender need not keep a retransmission list for each neighbor since PIM-Grafts are only sent to the RPF neighbor. Only the (S,G) entry needs to be tagged for retransmission.

2.1.3 Protocol scenario

A multicast datagram is sent by a source host. If a receiving router has no forwarding cache state for the source sending to group G, it creates an (S,G) entry. The incoming interface for (S,G) is determined by doing an RPF lookup in the unicast routing table. The (S,G) outgoing interface list contains interfaces that have PIM routers present and that do not violate the scoping limits of the group; the list also includes interfaces with host members for group G.

PIM-Prune messages received on a point to point link are not delayed before processing as they are in the LAN procedure. If the prune is received on an interface that is in the outgoing interface list, it is deleted immediately. Otherwise the prune is ignored.

When a multicast datagram is received on the incorrect LAN interface (i.e., not the RPF interface) the packet is silently discarded. If it is received on an incorrect point-to-point interface, prunes may be sent in a rate-limited fashion. Prunes may also be rate-limited on point-to-point interfaces when a multicast datagram is received for a entry with empty outgoing interface list.

2.2 PIM-Sparse Mode (PIM-SM)

Sparse-mode PIM operates by forwarding multicast data packets only on interfaces from which explicit join messages have been received. Receivers' designated routers (DR) send join messages to the RP for each active group. [1] Senders' designated routers send register messages to the RP, which in turn sends join messages up towards the source. Once the join messages have propagated upstream from the RP, data packets from the source will follow the (S,G) distribution path state established. The packets will travel to the receivers via the distribution paths established by the join messages sent upstream from receivers towards the RP. Multicast packets will arrive at some receivers before reaching the RP if the receivers and the source are both "upstream" of the RP. When the receivers initiate shortest-path distribution trees, additional outgoing interfaces will be added to the (S,G) entry; and RP-bit state is set up on the RP tree for that source. The data packets will be delivered via the shortest paths to receivers. Data packets will continue to travel from the source to the RP(s) in order to reach new receivers. Similarly, receivers continue to receive some data packets via the RP tree in order to pick up new senders. However, when source-specific(shortest-path) tree distribution is used, most data packets will arrive at receivers over a shortest path distribution tree.

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 implementers perspective, i.e., the actions performed by a single PIM router.

2.2.1 Local hosts joining a group

A host sends an IGMP Host-Report message identifying a particular group, G, in response to a directly-connected router's IGMP Host-Query message, as shown in 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 RP(s) for the group, G, see [2].

When a *designated router* (DR, see section 2.3.1) receives a report for a new group, G, the DR classifies

[1] The DR will assume the role of last-hop router for the receivers and send join messages to the RP. it might lose to other router by assert process later and then the DR is no longer responsible for sending join messages, see section 2.3.2.

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

the group as either wide area (RP-based) or not[2]; and if the group is RP-based the DR looks up the associated RP mapping. [3] A DR will identify a new group (i.e., one for which it has no existing multicast entries) as needing PIM-SM support by checking if there exists an RP mapping. If there is no RP mapping provided in IGMP RP-Report messages, and there is no mapping provided in the appropriate configuration file, then the router will assume that the group is to be supported with PIM-Dense Mode.

For the remainder of this description we will assume a single RP just for the sake of clarity. We discuss the direct extensibility to operation with multiple RPs later in the document in section 2.2.7.

The DR (e.g., router A in figure 1) creates a multicast forwarding cache for (*,G) . The RP address is included in a special record in the forwarding entry, so that it will be included in upstream join 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. A 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 sets a timer for this entry. The timer is reset each time an RP-Reachability message is received for (*,G), see section 2.2.2.

2.2.2 Establishing the RP-rooted shared tree

The last-hop router creates a PIM-Join/Prune message with the RP address in its join list with 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 is propagated along the RP tree. The WC-bit indicates that the address is an RP and the receiver expects to receive packets from new sources via this (shared tree) path, therefore upstream routers should create or add to (*,G) forwarding entries.

[2] For the remainder of this document we assume that the multicast address space is divided in such a way that this determination is made easily and consistently by routers.

[3] We have proposed the use of a new host IGMP RP-Report message that would allow hosts to inform their directly-connected PIM routers of G, RP(s) mappings. Hosts will learn of RPs in the same way they learn of multicast group addresses.

Figure 2: Example: Switching from shared tree to shortest path tree
Actions are numbered in the order they occur

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

The RP recognizes its own address and does not attempt to send join messages for this entry upstream. The incoming interface in the RP's (*,G) entry is set to null. RP-Reachability messages are generated by RPs 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 triggers joining towards an alternate RP, see section 2.2.5.

2.2.3 Switching from shared tree (RP tree) to shortest path tree (SPT)

When a PIM router has directly-connected members it first joins the RP tree. The router can switch to the sources' shortest path trees as soon as it starts receiving data packets from the sources. To do so the router detects data packets for G that are *not* sourced by an address Sn for which it has a multicast forwarding entry (Sn,G). As shown in figure 2, router A initiates a new multicast forwarding entry for (Sn,G), with an SPT-bit cleared indicating that the shortest path tree branch from Sn has not been completely setup, and in the mean time it still uses the shared tree to get packets from Sn. A timer is set for the (Sn,G) entry and this timer is reset whenever a data packet for (S,G) is received. [4] Only routers with local members initiate switching to the SPT; intermediate routers do not.

A PIM-Join/Prune message will be sent upstream router towards the new source, Sn, with Sn in the join list. The payload contains Multicast-Address=G, PIM-Join={Sn}, PIM-Prune=NULL.

When the (Sn,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 Sn arrives and matches on this entry, all receivers will continue to receive source packets along this path unless and until the receivers choose to prune themselves.

[4] This timer is also used in dense-mode PIM.

Note that a last-hop router may adopt a policy of not setting up a (S,G) entry (and therefore not sending a PIM-Join message towards the source) until it has received m data packets from the source within some interval of n seconds. This would eliminate the overhead of (S,G) state upstream when small numbers of packets are sent sporadically. However, data packets distributed in this manner may be delivered over the suboptimal paths of the shared RP tree. [5] The last-hop router may also choose to remain on the RP-distribution tree indefinitely instead of moving to the shortest path tree.

When a router with a (Sn,G) entry and a cleared SPT-bit starts to receive packets from the new source Sn on the interface used to reach Sn, it sets the SPT-bit, and sends a PIM-Prune towards RP, if its shared tree incoming interface differs from its shortest path tree incoming interface. This indicates that it no longer wants to receive packets from Sn via RP. In the PIM message sent towards the RP, it includes Sn in the prune list, with the RP-bit set indicating that an RP-bit state should be set up on the way to the RP [6]. The PIM-Join/Prune message payload contains Multicast-Address=G, PIM-Join=NULL, PIM-Prune={Sn,RPbit}. [7]

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), the RP-bit state has to be deleted upstream of the router; so as to bring all sources packets down to the new member.

In particular, the router should modify the local RP-bit state so that all sources' packets are sent down the arriving link for the join, but are notent down other previously-pruned branches. The router must trigger an (*,G) join upstream to eradicate RP-bit state upstream. If the arriving (*,G) join has a prune list in it, then those corresponding RP-bit entries should not need to be eradicated upstream.

2.2.4 Steady state maintenance of router state

In the steady state each router sends periodic refreshes of PIM messages upstream to each of the next hop routers that is en route to each source, S, for which it has a multicast forwarding entry (S,G); as well as for the RP listed in the (*,G) entry. These messages are sent periodically to capture state, topology, and membership changes. A PIM message is also sent on an event-triggered basis each time a new forwarding entry is established for some new (Sn,G) (note that some damping function may be applied, e.g., a merge time). Optionally the PIM message could contain only the incremental information about the new source. The delivery of PIM-Join/Prune messages does not depend on positive acknowledgment; routers recover from lost packets at the next periodic transmission.

2.2.5 Local hosts sending to a group

When a host sends a multicast packet, the DR must deliver it on the RP tree. This is done by the DR sending a PIM-Register packet to all known RPs. The data packet is encapsulated in the PIM-Register packet so the RP can deliver it to downstream members. The register informs the RP of a new source which causes it to send PIM-Join messages back to the source so all routers capture state. The routers

[5] Note that (S,G) state must be maintained in all last-hop routers when an SPT is maintained (and this is suboptimal when (*,G) and (S,G) overlap because you need both pieces of state to keep the joins going to the right places upstream).

[6] An RP-bit entry is a (S,G) entry on the RP tree. The RP-bit is set, indicating that the associated prune messages should be sent up the shared tree towards the RP, and that (S,G) joins should not be sent towards S. In addition, the outgoing interface from which it receives a PIM-Join/Prune message with (S,G) and the RP-bit in the prune list, is deleted from the outgoing interface list. Data packets matching the RP-bit state are not sent to that interface.

[7] Note that if the upstream interfaces of (S,G) and (*,G) of the router are the same LAN, then the next packet to arrive on the RP tree after the SPT tree join was sent will cause the SPT-bit to be set even though the packet came via the RP tree; because the router cannot distinguish between the previous hop router for data packets without looking at the Data Link address. If the RP tree previous hop is not the same as the shortest path previous hop, then the router will prune off of the RP tree. Consequently, if the RP is significantly closer to the receiver than the Source is, or if the Source join is lost and the RP tree prune is not, there may be a period of lost packets.

between the source and the RP maintain (S,G) state so they know how to get packets for source S to the RP. The DR can stop encapsulating data packets in PIM-Registers when it receives PIM-Register-Stop messages from the RPs.

If an RP has gone down during the register process, we want to limit how long we encapsulate data packets. Also, after the encapsulating stops and data is sent natively to the RP, it is desirable to know if the RP is still up. Therefore, there is a RP (liveness) timer, and an RP-status flag, kept per RP for all active groups in the DR of each source. The RP-timer is reset, and the RP-status flag is set to "up" when a PIM-Register-Stop message is received. When the RP-timer expires (for example, 270 seconds), an RP-status flag is set for that RP indicating that it is in a "down" state. The RP-status flag is initialized to "unknown".

The source's DR sends periodic Register messages with null data to the RP (for example, every 30 seconds) if it has not received any PIM-Register-Stop messages. The DR resets the RP-timer each time it receives a PIM-Register-Stop message. When the RP timer expires, an RP-status flag is set for that RP indicating that it is in a "down" state. Null-data Register messages continue to be sent it is determined when the RP comes back up. The RP will process the null-data Register message, and send a PIM-Register-Stop to the source router of the Register message.

When the DR detects that the RP has come back up (i.e. the RP status was "down" and it received a Register-Stop message), it flags each (S,G) that it is responsible for sending Registers for and changes the RP status to "up". When data arrives from any of those sources, Register messages encapsulated with data are sent to the RP so the RP can send joins back to the source to recapture state between the source and the RP.

2.2.6 Multicast data packet processing

Data packets are processed in a manner similar to existing multicast schemes. A router first performs a longest match on existing forwarding states based on the source and group address in the data packet. A (S,G) state will be matched first if there is one, otherwise an (*,G) state will be matched. 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 (whose timers have not expired). There are two exception actions that are introduced if packets are to be delivered continuously, even during the transition from a shared to shortest path tree. First, when a data packet matches on an (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 never trigger prunes. Data packets may trigger actions which in turn trigger prunes. For example, router *B* in figure 2 decides to switch to SPT at step 3, it creates a (Sn,G) entry with SPT-bit set to 0. When data packets from Sn arrive at interface 2 of *B*, *B* sets the SPT-bit to 1, which in turn triggers the sending of prunes towards the RP.

2.2.7 Multiple Rendezvous Points (RPs) and RP failure scenarios

If there is one RP then there is no concern about sources and receivers actually being able to rendezvous, but there is a single point of failure.

When multiple RPs are used, each source registers and sends data packets towards each of the RPs, but receivers only join towards a single RP. If one of the RPs fails, receivers that joined to that RP will stop receiving RP-Reachability messages and will start sending joins to one of the alternative RPs. Sources do not need to take special action.

Sender's DR keeps an RP-timer and RP-status flag per RP. Register messages must be sent to all RPs because there may have been last-hop routers that joined to different RPs. DR sends periodic Register messages (with null data) to the RP. The router resets the RP-timer each time it receives a PIM-Register-Stop message. When the RP timer expires, an RP-status flag is set for that RP indicating that it is in a "down" state. The DR checks the RP-status when it receives a Register-Stop. If the RP-status is down or unknown, DR sets the Register-bit in a bitmap for that RP in every (S,G) entry that uses that RP. The router also resets the RP-status flag to "up". The setting of the Register-bits causes data from the affected sources to be encapsulated in PIM-Register messages (again) and sent to that RP.

Unreachable RPs are detected by downstream routers using the RP-Reachability message. When a (*,G) entry is established by a router with local members, an (*,G) timer and an RP-status flag per available RP are set. The timer is reset each time an RP-Reachability message is received and The RP-status flag is initialized to "unknown". If this timer expires (for example, 270 seconds), the last-hop router looks up an alternate RP for the group, sends a join towards the new RP. The router modifies the incoming interface of the (*,G) entry to that used to reach the new RP. The outgoing interface list includes only those SM interfaces on which IGMP Host-Reports or PIM-Joins for the group were received. The router also sets an RP-status-timer; when this timer expires (for example, 90 seconds), the RP-status flag is reset to "unknown" to indicate that the router should be considered as a candidate (it is potentially up).

2.3 PIM-DM and PIM-SM Operation over Multi-access Networks

2.3.1 Designated router election

When there are multiple PIM routers connected to a multi-access LAN, 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 joins toward the RP and keeps track of all RPs' status for local senders.

A simple designated router (DR) election mechanism is used for both SM and DM PIM. 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.

DR election is only necessary on multi-access networks.

2.3.2 Parallel paths to a source or the RP

Two or more routers may receive the same multicast datagram that was replicated upstream. In particular, if two routers have equal cost paths to a source and are connected on a common multi-access network, duplicate datagrams will travel downstream onto the LAN. PIM will detect such a situation and will not let it persist.

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 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 toend 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 an PIM-Assert packet on the 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. [8] This is important so downstream routers send subsequent PIM-Joins/Prunes or PIM-Grafts to the correct neighbor.

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 LAN.

Asserts are also needed for (*,G) entries since there may be parallel paths from the RP and sources to a LAN. When an assert is sent for an (*,G) entry, the first bit (RP-bit) in the metric preference is always set to 1 to indicate that this path corresponds to the RP tree. So a SPT path will always look better than an RP-tree path.

Note that for a leaf LAN on the RP tree, it is possible that the DR will send joins to the RP and that packets will come down the RP tree through that DR even though it (the DR) is not the optimal path to the RP. We think that this is a reasonable situation given that RP trees do not provide optimal paths to begin with.

The DR may lose to another router on the LAN by the Assert process if there are multiple RP-tree paths traveling 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. If more than one RP tree paths travel through a particular LAN, RP-Reachability messages will make downstream routers merge to a single RP; no assert process is needed.

2.3.3 Join suppression

If a PIM-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 joins. A timer is set for the Joiner-bit; after it expires the Joiner-bit is set indicating further periodic joins should be sent for this entry. The Joiner-bit timer is reset each time a PIM-Join message is received from a higher-IP-addressed PIM neighbor.

2.4 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 or graft from downstream. Joins and grafts received on the current incoming interface are ignored. Joins and grafts 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 messages.

The PIM router must send a PIM-Join or PIM-Graft message out its new interface to inform upstream routers that it expects multicast datagrams over the interface. It must send a PIM-Prune message out

[8] The downstream routers will change their upstream neighbor to the router that sent the last PIM-Assert message during the assert process.

the old interface, if the link is operational, to inform upstream routers that this part of the distribution tree is going away.

The SPT-bit is also cleared in order to receive data packets via the existing RP tree (if it is still operational) before the new shortest path has been established. To override previous RP-bit state prunes, a join should also be sent to the upstream neighbor of (*,G) if the incoming interface (*iif*) of (*,G) is different from the *iif* of (S,G).

2.5 Timers

Each (S,G) and (*,G) entry have 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. The timer of an (S,G) entry is reset whenever a data packet for (S,G) is received, the timer for an (*,G) entry is reset when a packet arrives on the RP-tree and when (*,G) Joins are received. The timer for an (S,G) RP-bit entry is reset whenever an (S,G) prune with RP-bit set is received. The timer expires after 3 times the refresh period, typically it is 3 minutes (because the Joins are sent every 1 minute).

A timer is maintained for each outgoing interface listed in each (S,G) or (*,G) entry. The timer is set when the interface is added. A DM outgoing interface of a DM group stays active in the list as long as there is no prune received and there are live PIM neighbors or directly-connected group members. A outgoing interface timer of a SM group is reset each time a PIM-Join message is received on that interface for that forwarding entry (i.e., (S,G) or (*,G)). [9]

When a timer expires, the corresponding outgoing interface is deleted from the outgoing interface list if the associated group is supported with SM (i.e., RP based); and it is added to the outgoing interface list if the associated group is supported with DM (i.e., does not use an RP). When the outgoing interface list is null, a prune message is sent upstream and the entry is deleted after 3 minutes. [10] During this time the entry is known as a negative cache entry at which a prune is triggered. Once the (S,G) is timed out, it can be recreated when the next multicast packet or join arrives. When a *,G entry is deleted, all associated S,G,RPbit entries are also deleted.

There are timers associated with an RP per group. When an RP-Reachability messages is received or a Register-Stop message is received the timer is updated. RP-Reachability messages contain the time-out period. The RP timer must be set to this value. For a DR that is upstream of the RP, receipt of Register-Stop messages causes it to update its RP timer to 270 seconds.

2.6 Sparse Mode/Dense Mode interaction

[11]

If a group has RPs associated with it[12], then all members in PIM regions will join the group using

[9] When a timer is reset for an outgoing interface listed in (*,G) entry, we should also reset the interface timers for all (S,G) entries which contain that interface in their outgoing interface list. Because some of the outgoing interfaces in (S,G) entry are copied from (*,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). If there are sources in the prune list of the (*,G) join, then the timers for this 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 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 (*,G) entry exists; otherwise, data packets will travel down both RP tree and SPT. It may not result in periodic duplicates (because of the RPF check), but it does waste a lot of network bandwidth.

[11] Previous versions of the specification referred to the configuring of individual interfaces as SM or DM. We have since found this to be unnecessary and complicating. Henceforth SM and DM refers to a global characteristic of the group, not to a characteristic of elements of the network or elemenst of the group.

[12] We assume that this is determinable from the address itself.

the PIM-SM protocol, and packets will only be forwarded onto interfaces on which explicit join messages have been received; even if the interfaces are configured as DM. [13]

2.7 PIM/Non-PIM Interaction

Editors note: This part of the design has not been flushed out and will be updated in the next version of the spec.

Routers that have both PIM and non-PIM interfaces are configured as PIM/non-PIM Border Routers (PIM-BRs). All PIM-BRs join a special multicast group. Members of this group conduct a Designated BR (DBR) election among themselves. If a BR finds that it is the largest numbered participant in the DBR election, it sends an IGMP Query to the multicast group consisting of all multicast routers in the non-PIM domain.

Members of this group and BRs with downstream members respond by sending IGMP Host-Report messages to the group; members of group and BRs with downstream members also listen to these reports and suppress sending reports for groups that have been reported by other routers. As a result, all BRs hear of all groups for which internal or downstream members exist. There is an RP-entry BR (RBR) election per group, in which a single BR is elected to join towards the RP for that group. The election is based on the router with the shortest path towards the highest numbered RP in the RP list for that group. Any ties are resolved in favor of the higher numbered BR. RBRs are group specific.

Data packets will follow the resulting (*,G) join state down to the elected RBR, and into the non-PIM region. If the non-PIM region is part of the source rooted shortest path tree, then the data packets will be forwarded through the non-PIM region according to its internal RPF rules, and will arrive at all exit PIM-BRs that do not prune themselves. From there the data will be forwarded down the remainder of the tree. Data packets will not be flooded through the non-PIM region if they arrive via the wrong incoming border router, with respect to that source. Therefore we need to introduce some additional mechanism to cause RP tree packets to be forwarded through the non-PIM region. In order for the non-PIM cloud to propagate an unencapsulated data packet ¿from the RP tree, to any internal members and to other PIM-BRs (which might have downstream members), the packet must be injected via the PIM-BR(s) that are the shortest path tree entry points from the packet source, S, to the routers inside the non-PIM region.

To achieve this, the RP tree entry PIM-RBR must get the data packet to the PIM-BR(s) that are on the shortest path from the source to any part of the non-PIM region. To do so, the PIM-RBR can encapsulate the packet with its own address as source and multicast the packet to the all-PIM-BRs multicast address; the IP-protocol field is set to BR-encapsulate. When a PIM-BR receives a BR-encapsulated packet, it conducts two checks. First, if the PIM-BR has an (*,G) entry whose incoming interface points to the non-PIM region and it does not have the (S,G) entry, the PIM-BR forwards it to the outgoing interfaces specified in (*,G). [14] Second, if the PIM-BR's shortest path to the packet source is via an external route and it does not have the (S,G) entry, the PIM-BR forwards the packet into the non-PIM region as if it were arriving from the source's shortest path tree.

The RBR elected for a group is responsible for doing the SPT switch (if the data traffic, or other configured information, calls for it). When an RBR receives packets from source S over the RP tree, and it wants to switch to the SPT, the router sends an (S,G) join message to the all-PIM-BRs group. Every BR that has an "external" shortest path towards the source (i.e., the shortest path towards the source points outside the non-PIM cloud), sends an (S,G) join upstream towards the source. The resulting join

[13] We investigated an alternative approach in which wide-area-groups' data is distributed over DM interfaces in a data-driven DM fashion. However, the scheme required encapsulation of all data packets traveling on the RP tree (in SM, as well as DM regions), and appeared more complex to understand and implement.

[14] Those outgoing interfaces listed in an (*,G) entry should only point out to the PIM region.

state will cause unencapsulated packets from S to G to travel down the source-rooted tree and arrive at the BR(s) that sent joins. These packets will be flooded into the non-PIM cloud and reach all possible receivers and transits according to the native multicast mechanism.

When a source inside of a non-PIM region sends to a non-local group, the arriving packet (for which no (S,G) entry exists, and for which RP information does exist) triggers the same election procedure as was described above. In short, the BR with the shortest path to the highest numbered RP for that group, sends register packets to all of the RPs (with the encapsulated data); ties are resolved in favor of the highest numbered BR. RPs send joins back to the BR that sent the register and consequently the sources packets will travel out of the non-PIM cloud via that BR and down to the RP and other downstream receivers, according to the (S,G) state.

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.

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, and holdtime for the neighbor in the PIM neighbor timer; at which time, the Designated Router (DR) for the interface can be computed. The highest IP addressed system is elected DR. Each query received causes the stored information to be overwritten.

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 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.

3.2.1 Sending Join/Prune Messages

PIM-Join/Prune messages are used to construct or tear down multicast forwarding state respectively. Join/Prune messages are sent hop by hop towards the indicated sources.

A join is sent to construct forwarding state or to undo prune state. Joins are sent towards known sources based on the (S,G) state stored in the multicast routing table. Joins are also sent towards the RP for active (*,G) state.

A prune is sent to undo join state when members for a group are no longer present on a multicast tree branch. These prunes are sent towards known sources associated with (S,G) entries. Prunes are also sent on the RP tree for a source when a router decides to move off the RP tree and onto the shortest path tree.

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 multicasted to all routers on multi-access networks with the target address set to the next hop router towards S or RP.

These Join/Prune messages will be sent periodically. Currently the period is set to 60 seconds.[15]

A router will send a periodic Join/Prune message to each distinct RPF neighbor for each (S,G) and (*,G) entry it has in its multicast routing table. 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:

- An RP address (with RP and WC bits set) is included in the join list of a periodic Join/Prune message under the following conditions:
  - The Join/Prune message is being sent to the RPF neighbor to the RP.
  - The RP is determined to be in Up state, and
  - 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
- A particular source address, S, is included in the join list with the RP and WC bits cleared under the following conditions:
  - The Join/Prune message is being sent to the RPF neighbor to S, and
  - There exists an active S,G entry with the RPbit cleared, and
  - The oif list in the S,G entry is not null.
- A particular source address, S, is included in the prune list with the RP and WC bits cleared under the following conditions:
  - The Join/Prune message is being sent to the RPF neighbor to S, and
  - There exists an active S,G entry with the RPbit cleared, and
  - The oif list in the S,G entry is null.
- A particular source address, S, is included in the prune list with the RP bit set and the WC bit cleared under the following conditions:
  - The Join/Prune message is being sent to the RPF neighbor toward the RP and their exists an S,G entry with the RPbit set, or
  - The Join/Prune message is being sent to the RPF neighbor toward the RP, there exists an S,G entry with the RPbit cleared, and the RPF neighbor toward S is different than the RPF neighbor toward the RP.

In addition to these periodic messages, the following events will trigger PIM-Join/Prune messages:

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

[15] 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.

2. Receipt of a PIM-Join message for an (S,G) pair (including (*,G)) for which there is no current forwarding entry or the outgoing interface list of (S,G) entry is null, will trigger building (S,G) or new *oif* state (if the incoming interface is SM), and this will in turn trigger a PIM-Join message towards S (or RP) with S (or RP with RP-bit and WC-bits set) in the join list.

3. Receipt of packet on the new created (S,G) entry, i.e., an entry with SPT-bit cleared, over the appropriate incoming interface, and there is an (*,G) entry, triggers

   (a) setting of the SPT-bit on (S,G) entry, and

   (b) sending a PIM-Prune message up the RP tree, i.e., towards the RP, with S address and the RP-bit set in the prune list.

4. When the outgoing interface list of (S,G) entry becomes null, indicating no more downstream receivers, a PIM-Prune with S address in the prune list is sent upstream.

5. When a PIM-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 LAN.)

6. Receipt of a *,G Join message where the specified RP address is of a higher value than the RP value associated with an existing *,G entry. This triggers an updating of the RP value, iif, and an associated *,G Join message is sent toward the new RP and a prune is sent toward the old RP

We do not trigger prunes onto interfaces for SM group based on data packets. Data packets that arrive on the wrong incoming SM interface are silently dropped. Data packets that arrive on the wrong DM point-to-point interface of an DM group trigger a prune.

3.2.2 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 PIM-Join/Prune message if this PIM-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 Si in the join list has RP-bit and WC-bit set, Si is an RP address. 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). furthermore,

      i. For each (Si,G) entries associated with group G, if Si is not included in the prune list, then interface I is added to its *oif* list and reset the timers for that interface in each affected entry. If the (Si,G) entry is an RP-bit entry and its *oif* list is the same as (*,G) *oif* list, then the RP-bit entry is deleted.

      ii. If the RP address Si is different from the RP address listed in the existing (*,G) forwarding entry and Si is greater than the listed RP entry value, then set RP entry to Si; otherwise leave the RP entry as is.

   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.

(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. [16] The outgoing interface is set to I, and the incoming interface is set to the interface used to send unicast packets to Si. If the interface used to reach Si is the same as the outgoing interface being built, this represents an error and the join should not be processed.

(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 message, but the RP-bit is set on the existing (Si,G) entry, the router clears the RP-bit on (Si,G) entry, recomputes the incoming interface towards Si for that (Si,G) entry, and sends a join 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.

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

(d) For each address Si in the prune list, if there is an existing (Si,G) forwarding entry, the router schedules a deletion of I from the list of outgoing interface list 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. If the router has a current (*,G) forwarding entry, and if an (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:

   - An (Si,G,RP-bit) entry is created if there exists a *,G entry, but there does not exist a (Si,G) entry. The outgoing interface list copied from the (*,G) entry, with the interface 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.
   - If there exists a (Si,G,RP-bit) entry, then the entry timer is reset.

2. If the received join does not indicate the router as its target, then if the join is for an (S,G) pair for which the router has an active (S,G) entry, and if the join arrived on the *iif* for that entry. The router compares the IP address of the generator of the join, 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 joins should be sent for this entry. The Joiner-bit timer is reset each time a PIM-Join message is received from a higher-IP-addressed PIM neighbor.

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

[16] The router creates a (S,G) entry and copies all outgoing interfaces 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.

3.3 Graft and Graft-Acks

Grafts are sent up DM interfaces for DM group that have been previously pruned from the distribution tree.

3.3.1 Sending PIM-Grafts and receiving PIM-Graft-Acks

When a router in a DM region hears of a member of a previously pruned group, the router sends a graft message towards known sources for that group (i.e., towards sources for which the router has existing (S,G) entries).

The PIM-Graft message is the only PIM message that uses a positive acknowledgment strategy. Senders of PIM-Graft messages unicast them to their upstream RPF neighbors. The sender periodically retransmits the PIM-Graft message for any (S,G) that has not been acknowledged. Note that the sender need not keep a retransmission list for each neighbor since PIM-Grafts are only sent to the RPF neighbor. Only the (S,G) entry needs to be tagged for retransmission.

3.3.2 Receiving PIM-Grafts and sending PIM-Graft-Acks

When a router receives a graft message, it adds the receiving interface to its *oif* list; unless the receiving interface is the *iif* for the entry, in which case it is dropped. The receiving router processes each (S,G) in the graft message and immediately acknowledges each (S,G) in a PIM-Graft-Ack message. This is relatively easy, since the receiver simply changes the IGMP code from PIM-Graft to PIM-Graft-Ack, computes the checksum, and unicasts the original packet back to the source.

3.4 Assert

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

3.4.1 Sending asserts

The following Assert rules are provided when a multicast packet is received on an outgoing interface:

1. Do unicast routing table lookup on source IP address from data packet, and send assert on interface for source IP address from data packet, include metric preference of routing protocol and metric from routing table lookup.

2. If route is not found, use metric preference of 0xffffffff and metric 0xffffffff.

When an assert is sent for an (*,G) entry, the first bit (RP-bit) in the metric preference is set to 1, indicating the data packet is routed down the RP tree.

3.4.2 Receiving asserts

When an assert is received on an outgoing interface, the router performs a longest match on the source and group address in the assert message (either an (S,G) entry or an (*,G) entry will be matched). If the interface that received the assert is in the *oif* list of the matched entry, then this assert is targeted for this router and is processed as follows:

1. Compare metric received in assert with the one you 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 your value, prune the interface. If the value is the same, compare IP addresses, if your address is less than the assert sender, prune the interface.

2. If you have won the election and there are directly connected members on the LAN, keep the interface in your outgoing interface list. You are the forwarder for the LAN.

3. If you have won the election but there are no directly connected members on the LAN, schedule to prune the interface. The LAN might be a stub LAN with no members (and no downstream routers). If no subsequent joins are received, delete the interface from the outgoing interface list. Otherwise keep the interface in your outgoing interface. You are 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.

When an assert is received on an incoming interface, the router performs a 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. If there is a match, 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 LAN. This will cause the upstream forwarder to cancel its delayed pruning of the interface.

3.5 RP-Reachable

When an (*,G) entry is established by a router with local members, a timer is set. The timer is reset each time an RP-Reachability message is received. If this timer expires, the router looks up an alternate RP for the group, sends a join towards the new RP. A new (*,G) entry is established with the incoming interface set to the interface used to reach the new RP. The outgoing interface list includes only those interfaces on which IGMP Host-Reports for the group were received. (Other outgoing interfaces may no longer be valid since the router in question may not be on the shortest path between the downstream branch and the new RP. If the router is on this shortest path as well, it will eventually receive an explicit join from that downstream branch as the last-hop routers take the same action).

When multiple RPs are used, each source registers and sends data packets towards each of the RPs, but receivers only join towards a single RP. If one of the RPs fails, receivers that joined to that RP will stop receiving RP-Reachability messages and will start sending joins to one of the alternative RPs. Sources do not need to take special action.

Because each receiver's directly connected router selects an RP independently, it is possible for routers on the same part of the distribution tree to specify different RPs while both are still available. This can lead to looping in some topologies. To avoid looping, RP address information carried in PIM-Join and RP-Reachability messages is examined to converge to a common RP (the larger numbered RP dominates).

3.5.1 Sending RP-Reachability messages

A router starts sending periodic RP-Reachability messages downstream when it receives a PIM-Join/Prune message with its own address and WC-bit and RP-bit set in the join list; and the incoming interface on its (*,G) entry is null [17]. The first condition is to make sure that it is an RP. The second condition

[17] This rule is needed when the current node was specified as an RP, but yielded to another RP in the multiple RP address arbitration process. If the other RP's address is bigger, the current entry will have *none* null incoming interface

is to make sure that only the *dominant* RP will send RP-Reachability messages, so the traffic can be minimized.

This obviates the need to do any kind of special configuration of RPs; any router can be an RP since RP behavior is triggered by the protocol itself. A router is responsible for initiating RP-Reachability messages to downstream nodes if it has an (*,G) entry with a null incoming interface.

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 224.0.0.2 class D address and the message content includes the RP and G.

3.5.2 Receiving RP-Reachability messages

When a router receives an RP-Reachability message for a group G it must compare the RP address listed in the message to the RP address listed in the current (*,G) RP entry. If the RP listed in the message is greater than the RP listed in the (*,G) RP entry, and if the next hop used to reach the listed RP is the same as the next hop used to reach the RP entry, then the router replaces its current RP entry with the RP address from the RP-Reachability message When a router receives an RP-Reachability message it does the following (assume that router X receives an RP-Reachability message of RP1 from incoming interface I):

1. Perform RPF check. If I is not the best next hop to RP1, drop this RP-Reachability message.
2. If the incoming interface of (*,G) entry is *not* null and *not* I, drop the RP-Reachability message.
3. If the incoming interface of (*,G) is I, compare RP1 with the address in RP entry, say RP2. If RP1 is larger than RP2, set RP entry to RP1 and propagate the RP-Reachability message downstream. If RP1 is the same as RP2, then reset the entry timer and propagate the RP-Reachability message downstream. Otherwise, drop the RP-Reachability message.
4. If the incoming interface of (*,G) is null and RP-bit and WC-bit is set then this router is currently acting as an RP for G. In this case, compare RP1 with X. If RP1 is larger than X, set RP entry to RP1, set the incoming interface to the RPF interface used to reach RP1. Also, propagate the RP-Reachability message downstream. Otherwise, if RP1 is less than X, drop the RP-Reachability message.

3.6 Register and Register-Stop

When a source first starts sending to a group its packets are encapsulated in PIM-Register messages and sent to the RP(s). RPs send join messages towards the source(s) and send Register-Stop messages once the tree between the RP and source has been built.

3.6.1 Sending Registers

When a DR receives a multicast data packet from a directly connected host for which it has no (S,G) entry, or an existing (S,G) with a register bitmap set, it sends a register message to the RP(s) for that group. The message indicates the group for which the source is registering.

The original data packet is encapsulated inside the register packets. The message is sent as a unicast packet to the RP(s); it is not processed by the intermediate routers. If there are multiple RPs associated with the multicast group, then the source sends a register message to each of them.

Subsequent data packets sent to the same group will trigger the same action until *each* RP has built a (S,G) path back to the source and has sent a PIM-Register-Stop message to the source telling it to stop the sending of PIM-Register messages for (S,G).

The router sending the register messages maintains a register bitmap per (S,G) entry, with each bit referring to one of the RPs for that group.

1. When all bits are cleared, the router knows that all RPs have sent Register-Stop messages and no more register messages need be sent.
2. When some bits are set, the router continues to send data packets encapsulated in register messages to the associated RP(s).
3. The router maintains a RP-timer, it assumes a particular RP to be unreachable and sets the associated RP-status to “down” after the RP-time expires.
4. Periodically the router sends a null-data register packet to each RPs to verify that it is up. So long as Register-Stops are still received in return, the RP is considered up. If a Register-Stop is not received, a RP-timer will expire after 270 seconds and the RP will be considered down.

3.6.2 Receiving Register Messages

When a router (i.e., the RP) receives a register message, the router

1. Decapsulates the data packet, and forwards it according to its local (S,G) or (*,G) forwarding entry.
2. If there is no (S,G) entry, the RP sets up an (S,G) forwarding entry with the outgoing interface list copied ¿from the (*,G) outgoing interface list, its SPT-bit is set to 0. The (S,G) entry is set up using the mask information, if provided, in the register message. A timer is set for the (S,G) entry and this timer is reset whenever a data packet for (S,G) is received. The (S,G) entry causes the RP to send a PIM-Join message for the indicated group towards the source of the register message. The PIM-Join message includes the source’s address in the join list.

3.6.3 Sending Register-Stops

When unencapsulated data arrives from (S,G), the RP knows that the distribution path has been built between it and S and it can tell the source’s router to stop sending registers. Register-Stops are (S,G) specific. The RP continues to send Register-Stop messages so long as register messages continue to arrive; however the RP should rate-limit the sending of these messages to allow time for (S,G) join to arrive at the first-hop router.

RPs also send Register-Stop messages in response to the periodic null-data register messages.

3.6.4 Receiving Register-Stops

The DR checks the RP-status when it receives a Register-Stop. If the RP-status is “down” or “unknown”, the DR sets the Register-bit in a bitmap for that RP in every (S,G) entry that uses that RP. The router also resets the RP-status flag to “up”. The setting of the Register-bits causes data from the affected sources to be encapsulated in PIM-Register messages again and sent to that RP. This will stop when the router receives the corresponding PIM-Register-Stops from that RP.

Otherwise, if the RP-status is “up”, the DR clears the bit in its register bitmap for the associated (S,G) entry (when the bit is not set the router does not send encapsulated packets in registers for that (S,G) to that RP). The DR also resets the RP-timer.

3.7 Multicast Data Packet Forwarding

Processing a multicast data packet involves two 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.

2. Do an RPF check based on the source address in the packet header and the *iif* specified in the forwarding entry.

The processing actions depend on whether the group has RPs associated with it or not. For shorthand we will refer to DM groups as those for which there are no RPs, and SM groups as those for which RP(s) are defined.

3.7.1 Forwarding data packets for SM groups

The following actions are taken based on the results of the state lookup and RPF check if the group is SM:

1. 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 RP.

   (c) If the source of the packet is a directly-connected host and the router is the DR on the LAN, check the register bitmap associated with the (S,G) entry. If any of the bits are set, then the router encapsulates the data packet in a register message and sends it to the corresponding RP(s).

   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 SPT, 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.

2. If the packet matches to an entry but did not arrive on 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 an (S,G) entry for which the SPT has not yet been completely established upstream.

3. 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 the LAN and knows of RP(s) associated with the destination group, G, then the DR checks the register bitmap associated with the local sender (if there is no such a register bitmap, a new register bitmap is created and associated with the RP list, all bits are set), the data packet is encapsulated in register message(s) and sent to the RP(s) whose associated bit in the bitmap is set.

4. 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.

3.7.2 Forwarding data packets for DM groups

If the data packet is addressed to a DM group:

1. If a matching entry is found and the incoming interface check passes, the packet is forwarded to the *oif* list for the entry. If the *oif* list for the entry is null, a prune message may be sent upstream.
2. If there is no matching entry found but the RPF check passed: An (S,G) entry is created with the *oif* list populated with all DM-configured outgoing interfaces and all interfaces with local members, excluding the incoming one.
3. If the RPF check fails, a prune message may be sent upstream, and the packet is dropped.

3.7.3 Data triggered switch to shortest path tree

If a packet is received for directly connected members of a SM group, then if the longest match is (*,G) and the router is in a mode to prefer shortest path tree delivery for this group, an (S,G) entry is created and a join 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.

3.8 PIM to Non-PIM Border Routers

Editors Note: This section is in more flux than some of the others.

Routers that have both PIM and non-PIM interfaces are configured as PIM/non-PIM Border Routers (PIM-BRs)

3.8.1 DBR election

All PIM-BRs join a special multicast group. All multicast routers (PIM and non-PIM) join a second group. [18]

If a BR finds that it is the largest numbered participant in the DBR election, it sends an IGMP Host-Query to a multicast group consisting of all multicast routers in the domain; DBRs are group specific.

3.8.2 Receiving IGMP Host-Query message sent to all multicast routers

When a multicast router receives an IGMP Host-Query message sent to the all-multicast-routers address, it responds by sending an IGMP Host-Report message to the same group.

3.8.3 Receiving IGMP Host-Report message sent to all multicast routers

When a multicast router hears an IGMP Host-Report for which it also has members, it suppresses sending reports for those groups.

When a PIM-BR hears an IGMP Host-Report, it creates an (*,G) entry if:

- It knows a G, RP(s) mapping, and

[18] If it is not possible to modify non-PIM routers at all, then wherever a packet is indicated as sent to the all-multicast-routers address, it must be sent twice, once to the all-PIM-BRs address and once to the existing all non-PIM-multicast routers group. In the worst case it is sent to all routers and dropped by non-implementing routers.

- Its shortest path to the RP is an external path (i.e., the next hop is a PIM router, not a non-PIM router), and
- Its shortest path to the RP is the shortest among all PIM-BRs (i.e., It is the RP entry BR).

3.8.4 Generating join/prune messages

A PIM-BR sends a join towards an RP on a PIM link for each group for which it has an active (*,G) entry. (This is standard PIM behavior)

3.8.5 Forwarding multicast data packets from PIM links into non-PIM region

When a PIM-BR receives a packet on a PIM link, it conducts the usual RPF check and lookup based on the Source and Group address found in the packet header. If the packet arrived on the *iif* found in the longest-matched entry for (S,G) :

1. If a (S,G) match is found then:
   (a) the packet is forwarded to the PIM and non-PIM *oif*s in the *oif* list.
   (b) If the source of the packet is a directly-connected host, check the register bitmap associated with the (S,G) entry. If any of the bits are set, then the router encapsulates the data packet in a register message and sends it to the corresponding RP(s).
   (c) If the SPT bit was cleared and there is a local (*,G) entry, the SPT bit is set to one and a (S,G) prune message with RP-bit set is sent upstream towards the RP.
   (d) The timer is reset for the (S,G) entry.
2. If the group is SM and the longest match entry is a (*,G), then the BR multicasts the BR-encapsulated packet to the all-PIM-BRs address and sends the unencapsulated packet to all PIM interfaces in the *oif* list.
3. If no (S,G) matching entry is found (i.e., if no entry is matched, or if (*,G) is matched) but the group is SM and the source of the data packet is a local, directly-connected host, the data packet is encapsulated in register message(s) and sent to the RP(s). A register bitmap is created and associated with the RP list, all bits are set. (If any of the registrations had completed, there would have been an (S,G) entry in existence).
4. If the group is SM, no matching entries are found, and the source is not a local host, the packet is dropped.
5. If the group is DM and there is no matching (S,G) entry, if the BR is the entry BR for S, (it has an external path to s), then an (S,G) entry is created and the *oif* list is populated with all DM-configured interfaces and all non-PIM interfaces, excluding the incoming one; the SPT-bit is cleared. The BR multicasts the BR-encapsulated packet to the all-PIM-BRs address and forwards the unencapsulated packet to all PIM-DM interfaces in the *oif* list.

If a packet is received on a non-PIM interface and the router has PIM interfaces, the packet is processed in the same way as data packets arriving on PIM interfaces, with respect to the PIM interfaces.

3.8.6 Receiving BR-encapsulated data packets on all-PIM-BRs address

If a data packet is received on the all-PIM-BRs address and IP-protocol is set to BR-encapsulate, the router does an RPF check on the inner header. If the RPF check points to a PIM hop, and not to inside the non-PIM cloud, then the inner header source and destination addresses are used to do a longest match lookup. If a matching (S,G) entry with SPT-bit set is not found, then the router will inject the decapsulated packet into the non-PIM region. (This gets all the BRs that have external shortest paths to the source to inject the packet into the non-PIM region).

Furthermore, if there is an (*,G) entry whose *iif* is pointing inside the non-PIM region, then the BR is a RP exit BR, an RP-encapsulated version of the packet is sent to all PIM interfaces in the *oif* list. checked.

3.8.7 Forwarding multicast data packets from non-PIM links onto PIM links

A PIM-BR receives a packet from a non-PIM interface, if the RPF check succeeds and there is a (S,G) entry, the BR forwards the packet to all PIM interfaces in the *oif* list.

4 Packet Formats

RFC-1112, see [3], 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 which multicast groups they are members of. 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. 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-Stop
3 = Join/Prune
4 = RP-Reachability
5 = Assert
6 = Graft
7 = Graft-Ack

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.

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver|                   Reserved                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

PIM Ver PIM Version number is 1.

Reserved Transmitted as zero, ignored on receipt.

4.1 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                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Reserved          |         Holdtime              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

%%Dino
%     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
%     | Mode  |       Reserved        |          Holdtime             |
%     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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

Reserved Transmitted as zero, ignored on receipt.

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

4.2 PIM-Register Message

It is sent by the Designated Router (DR) to RPs when a multicast packet needs to be transmitted on the RP-tree. Source IP address is set to any 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                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                     Multicast data packet                     ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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

Multicast data packet The original packet sent by the source. For periodic sending of registers, this part is null.

4.3 PIM-Register-Stop Message

It is sent by the RPs in acknowledge receipt of a register message. Source IP address is the address the register was addressed to. 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                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Group Address                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Source Address                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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

Group Address The group address from the register message.

Source Address IP host address of source from multicast data packet in register. This address is set to 0 when a null register is sent.

4.4 PIM-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                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Upstream Neighbor Address                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Reserved            |           Holdtime            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Reserved    | Maddr Length  |  Addr Length  |  Num groups   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Multicast Group Address-1                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Multicast Group Address-1 Mask                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Number of Join Sources    |    Number of Prune Sources    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
|                    Join Source Address-1                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               .                               |
|                               .                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Join Source Address-n                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Prune Source Address-1                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               .                               |
|                               .                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Prune Source Address-n                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               .                               |
|                               .                               |
|                               .                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Multicast Group Address-n                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Multicast Group Address-n Mask               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Number of Join Sources    |    Number of Prune Sources    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Join Source Address-1                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               .                               |
|                               .                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Join Source Address-n                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Prune Source Address-1                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               .                               |
|                               .                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Prune Source Address-n                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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

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.

Maddr Length The length in bytes of the encoded multicast addresses.

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

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

Multicast group address For IP, it is a 4-byte Class D address.

Multicast group address mask A bit mask used against the multicast group address. This is the method to describe a range of multicast addresses. If the multicast group address field describes a single group address, the value must be 255.255.255.255.

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 below.

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.

```
 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  |S|W|R| Mask Len  |       Source Address ...      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  ...  Source Address          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Reserved Transmitted as zero, ignored on receipt.

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 Joins 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 Mask length is 6 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.

Source Address The address length is indicated from the Addr Length field at the beginning of the header. For IP, the value is 4 octets. This address is either an RP address (WC bit = 1) or a source address (WC bit = 0). When it is a source address, it is coupled with the group address to make (S,G).

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

4.5 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. 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 an 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                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Group Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Group Address Mask                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          RP Address                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Reserved            |           Holdtime            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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

Group Address The group address the RP is associated with.

Group Address Mask A bit mask that allows to describe group ranges. Must be set to 255.255.255.255 when Group Address describes a single group address.

RP Address The rendezvous point IP address of the sender.

Reserved Transmitted as zero, ignored on receipt.

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

4.6 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                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Group Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Group Mask                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Source Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|R|                       Metric Preference                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              Metric                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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

Group Address The group address the data packet was addressed to that triggered the Assert.

Group Mask Describes a range of group addresses. Use 255.255.255.255 when describing a single group.

Source Address Source IP address from IP multicast datagram that triggers the Assert packet to be sent.

R RP bit is a 1 bit value. If the IP multicast datagram that triggers 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.

4.7 PIM-Graft Message

This message is sent by a downstream router to a neighboring upstream router to reinstate a previously pruned branch of a source tree. This is done for dense-mode groups only. The format is the same as a PIM-Join/Prune message.

4.8 PIM-Graft-Ack Message

Sent in response to a received Graft message. The Graft-Ack is only sent if the interface in which the Graft was received is not the incoming interface for the respective (S,G). This is done for dense-mode groups only. The format is the same as PIM-Join/Prune message.

5 Pseudocode

Editors Note: This section is still in progress.

6 Acknowledgments

Tony Ballardie, Scott Brim, Jon Crowcroft, Paul Francis and Lixia Zhang provided detailed comments on previous drafts. The authors of CBT 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, and L. Wei. Protocol independent multicast (pim) : Motivation and architecture. *Internet Draft*, November 1994.

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

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