Protocol Independent Multicast (PIM), Dense Mode Protocol Specification

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

Ahmed Helmy

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

Van Jacobson

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

Liming Wei

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

draft-ietf-idmr-PIM-DM-spec-04.ps

Sept 12, 1996

Status of This Memo

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

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

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

1 Introduction

This specification defines a multicast routing algorithm for multicast groups that are densely distributed across an internet. The protocol is unicast routing protocol independent. It is based on the PIM sparse-mode [Estrin96] and employs the same packet formats. This protocol is called dense-mode PIM. The design is based largely on foundational work by Deering [Deering91].

2 PIM-DM Protocol Overview

Dense-mode PIM uses Reverse Path Multicasting (RPM). RPM is a technique in which a multicast datagram is forwarded if the receiving interface is one used to forward unicast datagrams to the source of the datagram. The multicast datagram is then forwarded out all other interfaces. Dense-mode PIM builds source-based acyclic trees.

Dense-mode PIM is data driven, whereby it is assumed that all downstream systems want to receive multicast datagrams. For densely populated groups this is optimal. If some areas of the network do not have group members, dense-mode PIM will prune branches of the source-based tree. When group members leave the group, branches will also be pruned.

Unlike DVMRP [DVMRP] packets are forwarded on all outgoing interfaces (except the incoming) until pruning and truncation occurs. DVMRP makes use of parent-child data to reduce the number of outgoing interfaces used before pruning. In both protocols, once truncation occurs pruning state is maintained and packets are only forwarded onto outgoing interfaces that in fact reach downstream members.

We chose to accept additional overhead in favor of reduced dependency on the unicast routing protocol, and reduced overall protocol complexity.

Dense-mode PIM differs from sparse-mode PIM in two essential points: 1) there are no periodic joins transmitted, only explicit triggered grafts/prunes, and 2) there is no Rendezvous Point (RP).

3 Background

Reverse Path Broadcasting (RPB) is different from RPF because duplicate packets are avoided in the RPB that are sent in RPF. In general, the number of duplicates sent on a link can be as high as the number of routers directly connected to that link.

Reverse Path Multicasting (RPM) is different from RPF or RPB because pruning information is propagated upstream. Leaf routers must know that they are leaf routers so that in response to no IGMP reports for a group, those leaf routers know to initiate the prune process.

In DVMRP there are routing protocol dependencies for a) building a parent-child database so that duplicate packets can be eliminated, b) eliminating duplicate packets on multi-access LANs, and c) sending "split horizon with poison reverse" information to detect that a router is not a leaf router (if a router does not receive any poison reverse messages ¿from other routers on a multi-access LAN then that router acts as a leaf router for that LAN and knows to prune if there are not IGMP reports on that LAN for a group G).

Dense-mode PIM will accept some duplicate packets in order to avoid being routing protocol dependent and avoid building a child parent database.

We introduce a simple prune mechanism for reducing duplicates on multi-access LANs. We introduce a simple graft mechanism to reduce join latency on previously pruned branches of a source-based multicast tree.

We introduce an alternative leaf-router detection mechanism that does not rely on a specific unicast routing protocol mechanism such as split horizon with poison reverse.

These mechanisms are described below.

4 Protocol Description

4.1 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. As a result, even if the DR for that LAN does not hear any IGMP Reports for a group, the DR will know to continue to forward multicast data packets to that group, and NOT to send a prune message to its upstream neighbor.

Since dense-mode 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 leaf router detects that there are no members downstream when it is the only 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 Router-Hello 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 224.0.0.13. The router to process the prune will be indicated by inserting its address in the "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 interfaces for the (S,G) entry from the prune list. The suggested delay time before deletion should be greater than 3 seconds.

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 "Address" field may contain the address of the wrong upstream router. To avoid this situation, the "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 224.0.0.13. The router to process the join will be indicated by inserting its address in the "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 schedule time for the deletion request expires, it deletes the outgoing LAN interface from the (S,G) multicast forwarding entry.

Note that if the join message is lost, the deletion will occur and there will be a no data delivery for the amount of time the interface remains in Prune state. To reduce the probability of this occurrence, a router that overrides a prune may send multiple joins back-to-back or have a small delay between successive joins.

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

4.2 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 optionally, send a PIM-Graft message towards known sources. This allows join latency to be reduced below that 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 Report or PIM-Graft was received for all known (S,G). If the (S,G) entry was a negative cache entry, 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.

Any routers receiving the PIM-Graft message, uses the received interface as an incoming interface for any (S,G) entry, will not add the interface to the outgoing interface list.

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 neighbor processes each (S,G) and immediately acknowledges each (S,G) in a PIM-GraftAck message. This is relatively easy, since the receiver simply changes the IGMP code from Graft to Graft-Ack and unicasts the original packet back to the source. 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.

4.3 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 dense-mode configured interfaces that have PIM routers present or host members for group G.

A PIM-Prune message is triggered when an (S,G) entry is built with an empty outgoing interface list. This type of entry is called a negative cache entry. This can occur when a leaf router has no local members for group G or a prune message was received from a downstream router which causes the outgoing interface list to become NULL. PIM-Prune messages are never sent on LANs in response to a received multicast packet that is associated with a negative cache entry.

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 negative cache entry.

4.4 Designated Router election

The dense-mode PIM designated router (DR) election uses the same procedure as in sparse-mode PIM. A DR is necessary for each multi-access LAN so a single router sends IGMP Host-Query messages to solicit host group membership.

Each PIM router connected to a multi-access LAN should transmit PIM Router-Hello messages every 30 seconds onto the LAN to support DR election. The highest addressed router becomes the DR. The discovered PIM routers should be timed out after 90 seconds. If the DR goes down, a new DR is elected.

DR election is only necessary on multi-access networks. It is not required that PIM Hello messages be sent on point-to-point links.

4.5 Parallel paths to a source

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. Dense-mode 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 224.0.0.13 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).

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 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. This is important so downstream routers send subsequent Prunes or 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.

The following Assert rules are provided:

Multicast packet received on outgoing interface:

1. Do unicast routing table lookup on source IP address from data packet.

2. Send Assert on interface for source IP address from data packet, include metric preference of routing protocol and metric from routing table lookup.

3. If route is not found, Use metric preference of 0x7fffffff and metric 0xffffffff.

Asserts received on outgoing interface:

1. Compare metric received in Assert with the one you would have advertised in an Assert. 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.

Asserts received on incoming interface:

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 packets should be forwarded on the LAN. This will cause the upstream forwarder to cancel its delayed pruning of the interface.

4.6 Timing out multicast forwarding entries

Each (S,G) entry has timers associated with it. During this time source-based tree state is kept in the network.

There should be multiple timers set. One for the multicast routing entry itself and one for each interface in the outgoing interface list. The outgoing interface stays active in the list as long as there is multicast traffic for the entry or there is an explicit Graft received on the interface. If neither occurs the interface will be deleted from the list after 3 minutes, by default.

Once all interfaces in the outgoing interface list are not active, a timer should be set for the (S,G) entry. During this time the entry is known as a negative state entry at which a prune is triggered. Once the (S,G) entry times out, it can be recreated when the next multicast packet or join arrives.

4.7 Source address aggregation and Pruning

An (S,G) entry in the multicast routing table will contain a source address to be as specific as necessary depending where the router is in relation to a source. Close to the source, this will typically be a subnet number. Far from the source, it may be a network number or a supernet route. Prunes sent may be rather ineffective if the source being pruned is not specific enough.

For example, initially a multicast datagram may be flooded throughout an Autonomous System (AS). Within the AS, there is complete subnet information in the unicast routing tables of all routers. Once the datagram exits the AS, it is likely there are routers that don't have subnet information. If these routers send Prunes for aggregate sources, routers close to the source will not know where to reach the source since they have more specific information than what was provided in the Prune message. This results in traffic being sent further on a branch of the multicast tree than necessary.

The problem is fixed by sending source host specific prunes. However, to maintain proper scaling of routing information, each router along the path performs longest match lookups for the source specified in the Prune message. Therefore, they keep the level of aggregation that best suits thier position to the source in the topology.

A source host specific Prune is encoded by copying the instigating source IP address from the multicast datagram into the PIM message using a mask length of 32.

5 Packet Formats

This section describes the details of the packet formats for PIM control messages.

All PIM control messages have protocol number 103.

Basically, PIM messages are either unicast (e.g. Registers and Register-Stop), or multicast hop-by-hop to 'ALL-PIM-ROUTERS' group '224.0.0.13' (e.g. Join/Prune, Asserts, etc.).

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|PIM Ver| Type  | Addr length   |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

PIM Ver PIM Version number is 2.

Type Types for specific PIM messages. PIM Types are:

0 = Hello
1 = Register
2 = Register-Stop
3 = Join/Prune
4 = Bootstrap
5 = Assert
6 = Graft (used in PIM-DM only)
7 = Graft-Ack (used in PIM-DM only)
8 = Candidate-RP-Advertisement

Addr length Address length in bytes. Throughout this section this would indicate the number of bytes in the Address field of an address, including unicast and group addresses.

Checksum The checksum is the 16-bit one's complement of the one's complement sum of the entire PIM message, (excluding the data portion in the Register message). For computing the checksum, the checksum field is zeroed.

For all the packet format details, refer to the PIM sparse-mode specification.

5.1 PIM-Hello Message

It is sent periodically by PIM routers on all interfaces.

5.2 PIM-SM-Register Message

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

5.3 PIM-SM-Register-Stop Message

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

5.4 Join/Prune Message

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

5.5 PIM-SM-Bootstrap Message

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

5.6 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.

5.7 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 Join/Prune message.

5.8 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 Join/Prune message.

5.9 Candidate-RP-Advertisement

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

6 References

[Deering91] S.E. Deering. Multicast Routing in a Datagram Internetwork. PhD thesis, Electrical Engineering Dept., Stanford University, December 1991.

[DVMRP] RFC 1075, Distance Vector Multicast Routing Protocol. Waitzman, D., Partridge, C., Deering, S.E, November 1988

[Estrin96] Protocol Independent Multicast Sparse-Mode (PIM-SM): Protocol Specification. D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, G. Liu, P. Sharma, L. Wei, September 1996

[Deering94b] An Architecture for Wide-Area Multicast Routing, S. Deering, D. Estrin, D. Farinacci, V. Jacobson, G. Liu,L. Wei, USC Technical Report, available from authors, Feburary 1994.

[RFC1112] Host Extensions for IP Multicasting, Network Working Group, RFC 1112, S. Deering, August 1989