Network Working Group
Internet Draft

Stephen Deering (cisco)
Deborah Estrin (USC)
Dino Farinacci (cisco)
Van Jacobson (LBL)
Ahmed Helmy (USC)
David Meyer (cisco)
Liming Wei (cisco)

draft-ietf-pim-v2-dm-01.txt

Nov 3, 1998

Protocol Independent Multicast Version 2 Dense Mode Specification

1 Introduction

This specification defines a multicast routing algorithm efficient for multicast groups that are densely distributed across a network. This protocol does not have a topology discovery mechanism often used by a unicast routing protocol. It employes the same packet formats sparse-mode PIM [PIMSM] uses. This protocol is called dense-mode PIM. The foundation of this design was largely built on Deering's early work on IP multicast routing [Deering91].

2 Terminology

The key words ``MUST'', ``MUST NOT'', ``REQUIRED'', ``SHALL'', ``SHALL NOT'', ``SHOULD'', ``SHOULD NOT'', ``RECOMMENDED'', ``MAY'', and ``OPTIONAL'' in this document are to be interpreted as described in RFC 2119.

3 Glossary

Reverse Path Forwarding (RPF) / RPF interface / RPF lookup

RPF is a multicast forwarding mode where a data packet is accepted for forwarding if it is received on an interface used to reach the source in unicast. The interface passing this check is called the RPF interface.

An RPF lookup for a source returns the RPF interface and the next-hop information, as if a route lookup is done for the source on a unicast routing table.

Topology Discovery Mechanism/Protocol

This mechanism provides sufficient topological information for a router to determine whether a neighbor system is upstream with respect to each multicast source. Some topology discovery mechanism can also determine if a neighbor is downstream with respect to each multicast source. An existing unicast routing protocol or a clone of it is sometimes used for this purpose (such as in DVMRP[DVMRP]). When a multicast routing protocol has a topology discovery mechanism built-in, the topology discovery mechanism is sometimes referred to as the unicast routing part of the protocol (even though it is

not used for forwarding unicast packets).

4 PIM-DM Protocol Overview

Dense-mode PIM assumes that when a source starts sending, all downstream systems want to receive multicast datagrams. Initially, multicast datagrams are flooded to all areas of the network. If some areas of the network do not have group members, dense-mode PIM will prune off the forwarding branch by setting up prune state. The prune state has an associated timer, which on expiration will turn into forward state, allowing data to go down the branch previously in prune state.

The prune state contains source and group address information. When a new member appears in a pruned area, a router can ``graft'' toward the source for the group, turning the pruned branch into forward state.

The forwarding branches form a tree rooted at the source leading to all members of the group. This tree is called a source rooted tree.

The broadcast of datagrams followed by pruning of unwanted branches is often referred to as a broadcast-and-prune cycle, typical of dense mode protocols. The broadcast-and-prune mechanism in dense mode PIM uses a technique called reverse path forwarding (RPF), in which a multicast datagram is forwarded if the receiving interface is the one used to forward unicast datagrams to the source of the datagram.

Compared with multicast routing protocols with built-in topology discovery mechanisms (e.g. DVMRP with its own RIP-like ``unicast'' routing protocol), dense mode PIM has simplified design, and is not hard-wired into a specific type of topology discovery protocol. However, such simplification does incur more overhead and cause broadcast-and-prune to occur on some links that could be avoided if sufficient topology information is available, e.g. to decide whether each interface leads to any downstream neighbors for a particular source. We chose to accept the additional overhead in favor of the simplification and flexibility gained by not depending on a specific type of topology discovery protocol.

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

5 Protocol Description

Dense mode PIM initiates forwarding state in routers when a source begins to send. A source does not give any prior notifications to the network when it sends multicast datagrams to a group G. If a receiving router does not already have a forwarding entry, it creates it for the source and group G. This forwarding entry is called a (S,G) entry. It includes the following contents: source address, group address, the incoming interface, a list of outgoing interfaces, a few flags and a few timers. The incoming interface for (S,G) is determined by an RPF lookup in the unicast routing table. The (S,G) outgoing interface list contains interfaces that have PIM routers present or host members for group G.

If a router creates a (S,G) entry with an empty outgoing interface list after receiving a multicast datagram, it must trigger a PIM-Prune message toward the source S. This type of entry is called a negative cache entry. Negative cache entries can be found on leaf routers with no local group members, or on routers where prune messages were received from downstream routers that caused the outgoing interface list to become NULL.

Dense mode PIM routers send periodic Hello messages out of each interface and keep track of neighbors based on received Hello messages. The Hello message has a Holdtime field that tells the neighbor to delete neighbor information if it is not refreshed before expiration.

The following sections describe in detail how leaf network is defined and detected; how pruning is done on multi-access LANs; and actions related to new members joining an existing group; issues with designated router election; parallel path resolution; multicast forwarding entry expiration and the adaptation to topology changes.

5.1 Leaf network detection

In dense mode PIM, prune state is first instantiated on routers connected to leaf networks without group members. A network on a router interface is deemed a leaf if there is no other PIM neighbors on that network. The notion of leaf network here might be slightly different from that defined in other protocols. [*]

---

[*] E.g. In DVMRP a leaf network for a source is defined as one without downstream neighbor DVMRP routers. Each DVMRP router is able to decide whether it has a downstream DVMRP neighbor with respect to each source. This is due to the RIP-like mechanism used in its to-

A router determines it is a leaf by not receiving PIM Hello messages. A leaf router can detect 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. A router should not populate a forwarding entry's outgoing interface list with a leaf network interface without downstream members.

It should be noted that a non-leaf router in PIM dense-mode is not necessarily a non-leaf node on a particular multicast forwarding tree. There are topologies where two parallel routers are connected to a common LAN where none of the routers uses the other one to reach a source. Therefore both routers will be leaves of any forwarding trees rooted at the source. When there are no group members on this LAN, both routers must not forward packets onto it. This is achieved by an assert process. Please see section 5.5 for more details.

A dense mode PIM implementation must take proper actions when a non-leaf router becomes a leaf router. When the last PIM neighbor on an interface goes away and turns the connected subnet into a leaf network, the router should delete that interface from the outgoing interface lists of all groups without attached group members.

5.2 Pruning of branches

5.2.1 Pruning on multi-access LANs and prune-override

To avoid PIM-Prune message storms, PIM-Prune messages must not be sent on LANs in response to each received multicast packet that is associated with an existing negative cache entry.

A prune is sent upstream when a router creates a (S,G) entry with an empty outgoing interface list, or when the outgoing interface list becomes empty. [*]

Prune information is flushed periodically. This causes multicast datagrams to be sent in RPF mode again which in turn triggers prune messages.

When a prune message is sent onto an upstream LAN, it is data link multicast and IP addressed to the all PIM routers group address 224.0.0.13. The router to process the prune will be indicated by inclusion of its address in the "Address" field of the message. This address is obtained by an RPF look up for the source. When the prune

---

pology discovery (or ``unicast routing'') protocol.
[*] I.e. the set of forwarding interfaces in the out-going interface list is NULL.

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. The default prune delay time is 3 seconds.

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 PIM routers group address 224.0.0.13. The router to process the join will be indicated by inclusion of its address in the "Address" field of the message. The address is determined by an RPF lookup for the source. When the expected router receives the join message, it will cancel the deletion request. This process is called prune-override.

Routers that need to send prune-override joins 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 control traffic on the LAN. The maximum join delay timer should be equal to or smaller than the prune delay timer value. The default is 3 seconds.

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

However, if the prune-override join message is lost, the deletion will occur and there will be no data delivery for the amount of time the interface remains in prune state. To reduce the probability of this occurance, a router that has scheduled to prune the interface off is required to immediately send a prune onto the interface. This gives other routers another chance to hear the prune, and to respond with a join.

Additionally, equal-cost paths leading to a LAN presents a special case for join/prune processing. 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.

5.2.2 Pruning a Point-to-point link

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.

Prunes may be rate-limited on point-to-point interfaces when a multicast datagram is received for a negative cache entry, or if the point-to-point interface receiving the datagram is not the RPF interface toward the source.

5.3 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 toward known sources. This allows join latency to be reduced below that indicated by the relatively large timeout value suggested for prune information.

The host indicates its interest in group G by sending an IGMP report. 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)'s. If the (S,G) entry was a negative cache entry, the router sends a PIM-Graft message upstream toward S.

A router receiving the PIM-Graft message adds the received interface into the matching (S,G) entry's outgoing interface list. If the entry transitions to forward state due to this added outgoing interface, the router must send a PIM-Graft message toward the source.

Routers with no group state do nothing on receipt of IGMP reports, since dense-mode PIM will deliver multicast datagrams to all interfaces when creating state for a group.

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 PIM message type 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. The interval between each retransmission is 3 seconds. 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.

5.4 Designated Router Election

Dense mode PIM itself does not need the function of a designated router (DR). But a DR is needed on multi-access LANs running IGMP version 1, which relies on a routing protocol to select a query router for the purpose of sending IGMP Host-Query messages. Dense-mode PIM designated router (DR) election has the same procedure sparse-mode PIM uses.

Each PIM router connected to a multi-access LAN should periodicly transmit PIM-Hello messages onto the LAN. The period is specified as [Hello-Timer] in the sparse mode specification (default 30 seconds). The highest IP addressed router becomes the DR. The discovered PIM routers should be timed out after 105 seconds (the [Neighbor-Timer] in the PIM-SM specification]. If the DR goes down, a new DR is elected.

DR election is only necessary on multi-access networks.

5.5 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 an outgoing interface on a multi-access LAN, 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 (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 best metric to the source. 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. If metrics are comparable, the router with the best metric will become the forwarder. Incomparable metrics will be discussed later in this section when metric preference is described. 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 PIM routers on the LAN.

Asserts are rate-limited. The recommended minimum interval between two subsequent asserts out the same outgoing interface of an (S,G) entry is 1 second.

The following Assert rules must be observed:

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. The assert message includes 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, and 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.

4 The assert winner sends an assert with its own metric on the LAN, so that all other routers know who the winner is.

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.

If the upstream router selected is different from the RPF neighbor selected natively, it should start an Assert-Timer. When this timer expires, it restores the RPF information obtained by the RPF lookup.

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.

3 Whent he assert state times out, the downstream router may switch from the assert winner to the original RPF neighbor, if different.

5.6 Timers in Multicast Forwarding Entries

An (S,G) entry has a number of associated timers. One for the multicast routing entry itself, one for each pruned interface in the outgoing interface list and one to time out information about upstream assert winner (Assert_timer). An outgoing interface in forward state does not time out or change state without external events. The outgoing interface stays in forward state in the list as long as there is a group member connected, or there is a downstream PIM neighbor that has not sent a prune to it. The interface is deleted from the outgoing interface list if it is on a leaf network and there is no connected member. The interface timer for a pruned interface should be started with the holdtime in the prune message (also referred to as

the prune timer).

When a (S,G) entry is in forwarding state, its expiration timer is set to [Data-Timeout], as specified in the companion sparse mode specification [PIMSM], which is 210 seconds by default. This timer is restarted when a data packet is being forwarded. The (S,G) entry is deleted if this timer expires.

Once all interfaces in the outgoing interface list are pruned, the (S,G)'s expiration timer should be set to the maximum prune timer among all its outgoing interfaces. During this time the entry is known as a negative cache entry.

If the prune timers on different outgoing interfaces are different, the pruned interface with the shortest remaining timer may expire first and turn the negative cache entry into forwarding state. When this happens, a graft should be triggered upstream. When a negative cache entry turns into a forward entry, the entry timer should be restarted with the default expiration timer. Once the (S,G) entry times out, it will be recreated when the next multicast packet or join arrives.

The prune message sent upstream contains a holdtime. Its default value is [Data-Timeout], except when the Assert timer is also running, the holdtime in the prune message is set to the smaller value between the prune holdtime the system uses, and the remaining assert timer value before its expiration. The Assert Timer is started with a default value of 210 seconds.

5.7 Adapting to unicast route changes

When unicast route changes occur, the RPF interface for many (S,G) entries will also change. The following should be done for the affected entries:

1 A prune should be sent toward the old incoming interface;

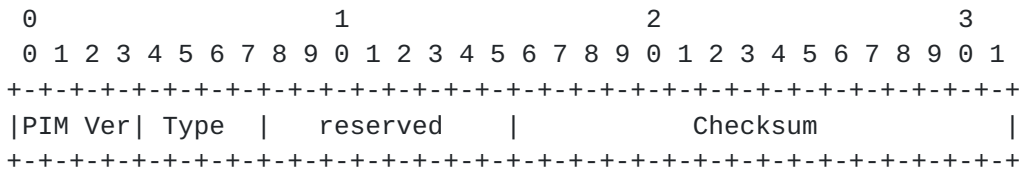
2 A graft should be sent to the new RPF neighbor.

6 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.).



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

Reserved
Set to zero. Ignored upon receipt.

Checksum
The checksum is the 16-bit one's complement of the one's

complement sum of the entire PIM message. For computing the checksum, the checksum field is zeroed.

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

6.1 PIM-Hello Message

It is sent periodically by PIM routers on all interfaces.

6.2 PIM-SM-Register Message

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

6.3 PIM-SM-Register-Stop Message

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

6.4 Join/Prune Message

It is sent by routers toward upstream sources. A join creates forwarding state and a prune destroys forwarding state. Refer to PIM sparse-mode specification.

6.5 PIM-SM-Bootstrap Message

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

6.6 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. This message is used in both dense-mode and sparse-mode PIM. For packet format, refer to PIM sparse-mode specification.

6.7 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, except that the value in the type field is 6. The source address hsould be included in teh join section of the message. The holdtime field is unused, and is ignored when a graft is received.

6.8 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, except that the value of the message type field is 7. The ``Encoded-Unicast-Upstream Neighbor Address'' field is unused and needs not be checked when this message is received. The holdtime field in the packet is ignored when received.

7 Acknowledgement

Thanks to Manoj Leelanivas, Sangeeta Mukherjee, Nitin Jain and many members of the PIM/IDMR working group for their helpful comments.

8 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

[PIMSM] Protocol Independent Multicast Sparse-Mode (PIM-SM): Protocol Specification. D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P.Sharma, L. Wei draft-ietf-idmr-pim-sm-specv2-00.txt.

[PIMARCH] 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, Nov 1996.

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

9 Security Considerations

Security issues are not discussed in this memo.

10 Authors' Addresses

Stephen Deering
Cisco Systems Inc
170 West Tasman Drive,
San Jose, CA 95134
deering@cisco.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

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

David Meyer
cisco Systems, Inc
170 West Tasman Drive
San Jose, CA95134
meyer@cisco.com

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