IDMR Working Group
INTERNET-DRAFT
Expires April 1993
<draft-ietf-idmr-igmp-dense-00.txt>

Steve Deering
Xerox PARC
Deborah Estrin
USC/ISI
Dino Farinacci
cisco Systems
Van Jacobsen
LBL

October 10, 1993

IGMP Router Extensions for Routing to Dense Multicast-Groups

Status of this Memo

1.0 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 ESL sparse-mode [Estrin93] and employs the same packet formats. This protocol is called dense-mode ESL. The design is based largely on foundational work by Deering [Deering91].

2.0. Overview

Dense-mode ESL 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 ESL builds source-based acyclic trees.

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

3.0. Background

Reverse Path Broadcasting (RPB) is different from RPF because duplicate packets are avoided in the former that are sent in the latter. 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 ESL 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 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.

4.0 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 ESL 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 ESL Router-Query 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 (224.0.0.1). 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.

Note the special case for equal-cost paths. When an upstream router is chosen by an RPF lookup there may be equal-cost paths. 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 ESL-Join message is data link multicast and IP addressed to the all routers group address (224.0.0.1). 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.

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

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, multicast a ESL-Join 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 an entry for (S,G), it adds the interface on which the IGMP Report or ESL-Join was received. If the (S,G) entry was a negative cache entry, the router sends an ESL-Join upstream towards S. This is done for all (Si,G) entries.

If routers have no state for (*,G), they do nothing since dense-mode ESL will deliver a multicast datagram to all interfaces when creating state about a group.

Any routers receiving the ESL-Join that uses the received interface as an incoming interface for any (Si,G) entry, will not add the interface to the outgoing interface list.

The ESL-Join message is transmitted unreliably.

4.3 Protocol Scenario

A multicast datagram is sent by a source host. If a receiving router has no forwarding cache entry for G, it creates (S,G) and (*,G) entries. (*,G)->incoming = NULL and (*,G)->outgoing set to all other interfaces on the router that have either multicast hosts or routers present. (S,G)->incoming = interface from RPF lookup. (S,G)->outgoing is copied from (*,G)->outgoing minus the (S,G)->incoming.

An ESL-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. ESL-Prune messages are never sent in response to a received multicast packet that is associated with a negative cache entry.

ESL-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 it is ignored.

4.4 Designated Router election

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

Each ESL router connected to a multi-access LAN should transmit ESL Router-Query messages every 30 seconds onto the LAN to support DR election. The highest addressed router becomes the DR. The ESL routers discovered 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 ESL Query 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 ESL 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 the highest IP addressed system should be elected to be forwarder for this (S,G) entry. When such a datagram is received, it triggers an ESL-Assert message to be multicast to 224.0.0.2 on the LAN. Each router that uses the LAN as an outgoing interface for (S,G) will compare the source IP address from the message with its own. If its own address is smaller, it will delete the interface from the outgoing interface list for the (S,G) entry. Otherwise, it has been elected as forwarder and will keep the interface in the entry.

This mechanism assures that only one router will forward multicast datagrams from S to G onto the LAN.

Interfaces that are pruned due to Assert processing should have a shorter timer associated with it compared to the timer that is used when an interface is pruned due to receipt of ESL-Prune message. This is recommended so unicast routing changes upstream do not cause long lived black holes.

Assert messages are data link multicast and IP addressed to the all routers group address 224.0.0.1.

4.6 Timing out multicast forwarding entries

Each (S,G) and (*,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 join received on the interface. If neither occurs the interface will be deleted from the list after 90 seconds, 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.

5.0 Sparse mode compatibility

There are two issues to consider when dealing with dense-mode ESL and sparse-mode ESL interaction.

- Is a group part dense and part sparse (dual-mode).
- If a group is either dense-only or sparse-only, when should it transition to the other mode.

5.1 Dual-mode

If a group has membership qualities where it is densely populated in some areas of the network and sparsely populated in others, it will have to be in both modes for the group. If a group has one or more RP addresses associated with it, either dynamically determined or through configuration, it will operate in sparse-mode. In addition, if any interface is configured as a dense-mode interface for the group, the router will also operate in dense-mode. The group is known to be in dual-mode.

In dual-mode, a (*,G) entry's outgoing interfaces are built from the union of dense-mode configured interfaces and interfaces where ESL-Join messages have been received. A (*,G) entry's incoming interface is always set to NULL. An (S,G) entry's outgoing interfaces are always copied from the (*,G) entry's outgoing interfaces. A (S,G) entry's incoming interface is based on the interface determined by an RPF lookup when a multicast packet is received for a (*,G) entry.

Outgoing interface entries are only timed out if they are not dense-mode configured interfaces. Typically, these are interfaces that were created from an sparse-mode ESL-Join. The interface is kept active by periodic receipts on ESL-Join messages from downstream routers.

If an ESL-Join is received on a dense-mode configured interface, it should be propagated upstream if there is no (S,G) or (*,G) entry. Otherwise, it is ignored.

Periodic ESL-Join and Prune messages are sent upstream if the interface is not configured in dense-mode. Otherwise, periodic messages are not sent.

A router that is upstream from the RP, will send ESL-Register messages to the RP if there is one configured for an associated group. The Register messages may travel on dense-mode configured interfaces.

5.2 Mode Transitioning - from one mode to another

When a router decides the mode for a group is not efficient, it may want to change to the alternate mode or dual-mode.

1) An upstream router can change a downstream interface from sparse to dense without informing downstream router. This change indicates that the upstream router doesn't require periodic Joins/Prunes or Registers to keep multicast cache entries from timing out. The downstream router can stop sending periodic messages when told that the mode has changed. However, this action is not required.

2) An upstream router needs to tell a downstream router when going from dense to sparse so the downstream router can start sending periodic messages.

An upstream router sends an ESL-Mode message to neighboring downstream routers indicating that it wants to change the mode for their common network. The mode attribute is for all (Si,G). A receiving router is required to acknowledge the ESL-Mode message with an ESL-ModeAck.

ESL-Mode messages are initially multicast and later retransmitted as unicast on multi-access LANs if an acknowledgement is not received.

Downstream routers that receive an ESL-Mode message switch their incoming interface to the mode indicated in the message. Upstream routers unilaterally control the mode.

Sparse-mode ESL RP-Reachable messages must be sent to all downstream interfaces regardless if they are in dense or sparse mode. This allows any downstream sparse-mode router to determine if the RP is still reachable. An RP-Reachable message is sent downstream in response to a received ESL-Join with the RP address in the source join list part of

the message.

6.0 Message Types

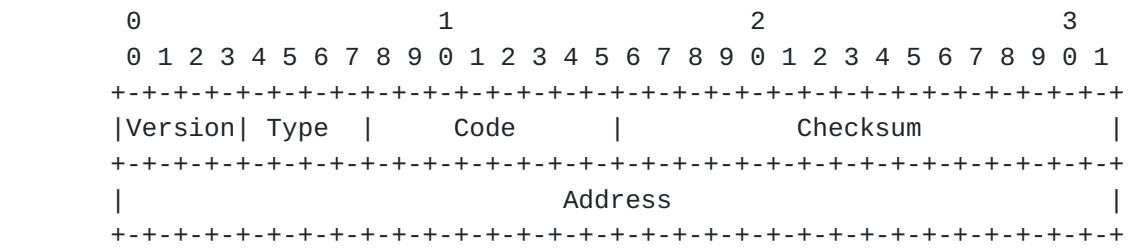
ESL messages are encapsulated in IGMP. IGMP runs on top of IP.

The following table enumerates the IGMP messages used in each mode.

	Sparse-mode	Dense-mode
(1) Host Membership Query	Sent	Sent
(2) Host Membership Report	-	-
(4) Router ESL		
(0) Query	Sent/Received	-
(1) Register	Sent/Received	-
(2) Join/Prune	Sent/Received*	Sent/Received**
(3) RP-Reachable	Sent/Received	-
(4) Assert	-	Sent/Received
(5) Mode	Sent/Received***	Sent/Received***
(6) ModeAck	Sent/Received***	Sent/Received***

* Sent periodically
** Sent only triggered by event
*** Sent and received only if Mode Transitioning is supported

6.1 Code field addition to IGMP packet format



A Code field has replaced the Unused field to support the variations of the Router ESL message. The Code should be set to 0 and ignored on receipt for all Types other than type 4.

Hosts must ignore the Code field.

Code values:

0 - Query

"Address" is set to 0 and ignored on receipt.

1 - Register

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

2 - Join/Prune

Regular sparse-mode/dense-mode Join/Prune list for adding or deleting branches to/from a source or RP-based distribution tree. The format is in the ESL sparse-mode specification.

When "Address" is used by dense-mode for leaf network detection, it contains the IP address of the router which processes the Join or Prune. Otherwise, it is set to 0 and ignored on receipt.

3 - RP-reachable

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

4 - Assert

This message is used for electing one of multiple parallel routers for downstream forwarding on a multi-access LAN. The body of this message is identical to the Join/Prune (2) format.

5 - Mode

This message is used by an upstream router to inform downstream neighbors its desireability to change modes. "Address" is set to the group address the mode is associated with.

6 - ModeAck

This message is sent by a downstream router to a neighboring upstream router that has previously sent an ESL-Mode message. This acknowledges the ESL-Mode message was received without error. "Address" is set to the group address the mode is associated with.

7.0 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

[Estrin93] IGMP Router Extensions for Routing to Spare Multicast-Groups, S. Deering, D. Estrin, D. Farinacci, V. Jacobson, September 1993

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

8.0 Interoperability Issues

1) MSOPF/ESL-DM interaction
2) DVMRP/ESL-DM interaction
3) ESL over NBMA links
   - Either still do BMA procedure with data link replication
   - (S,G) entries need to have neighbor IP addresses
   - Do we IP unicast or just data link unicast
