Network Working Group
INTERNET-DRAFT
Expiration Date: November 1999

Deborah Estrin
USC/ISI
Dino Farinacci
cisco Systems
May 17, 1999

Bi-Directional Shared Trees in PIM-SM

<draft-farinacci-bidir-pim-01.txt>

Status of this Memo

Abstract

This proposal extends the PIM-SM [1] mechanism for multicast datagram forwarding. PIM-SM constructs and maintains uni-directional shared trees and uni-directional source trees. We describe how we can extend the elements of operation of sparse-mode PIM to support bi-directional shared trees.

1. Introduction

A uni-directional shared tree allows sources to send multicast datagrams to members of a multicast group. Members receive packets sent to the group by joining the shared tree, using a particular node in the network as the root of the shared tree. The root of the shared

tree is called the Rendezvous Point (RP). When using undirectional shared trees, all sources' datagrams initially go to the root (RP) of the tree before being delivered down the distribution tree. As a result, there can be suboptimal delivery paths to the receivers close to the source.

In PIM-SM, the RP typically joins back to the source to draw datagrams down directly and natively (no encapsulation) from the source to the RP. The RP then forwards the datagrams down the uni-directional shared tree to the receivers. Eventually, receivers may join to the source as well, thus drawing datagrams down a source specific, uni-directional, shortest path tree. Or the receivers may continue to receive datagrams on the shared tree.

When using bi-directional shared trees, data can flow in either direction on a branch of the tree. This allows improved data delivery to receivers close to the source because the traffic traveling upstream to the root node is "turned around" and forwarded on downstream branches [2].

The bi-directional shared trees described in this extension to PIM-SM are used both to distribute datagrams from sources to the RP, as well as to distribute datagrams directly to receivers. Moreover, the protocol does not build source-specific trees from sources to the RP, nor to receivers. Instead, source transmissions travel up the shared tree toward the RP providing coverage to receivers along the way. The RP only needs to forward datagrams downward on those branches of the shared tree not covered by the path from the source to the RP.

However, bi-directional trees are incompatible with source specific uni-directional trees and so no switching to source-trees is allowed. Source-trees have the best delay characteristics so there is a tradeoff between uni-directional shared trees with source-trees and bi-directional shared trees. For large numbers of moderate to low rate sources, bi-directional PIM may offer significant advantages.

2. Pros and Cons of Bi-Directional Shared Trees

There are 3 basic advantages of bi-directional shared trees:

1. State is reduced compared to source trees. Each router in the multicast routing domain needs only keep state for the group and not each source sending to each group. [2]

2. Datagrams from sources to topologically near-by receivers do not have to travel all the way to the root of the shared tree. These improved distribution paths also support better scoping semantics for

applications that might use TTL based expanding ring scope to look for nearby resources.

3. Bursty sources can send with no or little state in routers.

There are 3 basic disadvantages of bi-directional shared trees:

1. Since all traffic eventually goes to the root of the tree, there is a traffic concentration point at the root node and links leading to it (pruning mechanisms could be added but at the cost of additional state and complexity). Traffic always flows to the root node even when it doesn't have to. That is, if the root node has a single sender branch, the root does not take part in forwarding traffic but it must receive the traffic because downstream nodes don't know the group membership tree near the root.

2. The path taken between the source and receivers might not travel over the shortest path, although it is likely to be a shorter path than via a uni-directional shared tree.

3. Bi-directional trees are incompatible with uni-directional source-trees. There is an increase in complexity when both are used for the same group.

Compared to CBT, the bi-directional trees proposed in this specification differ in two respects:

1. Non-member senders do not encapsulate their data to the root, the data is forwarded along the same path that it would take if the sender were also a member.
2. The protocol reuses much of the existing PIM-SM implementation.

3. Modifications to PIM

A strong goal of this proposal is to make as few changes as possible to PIM and multicast forwarding. We also wish to make the changes compatible, to enable a phased (incrementally deployed) transition to bi-directional shared tree PIM. Therefore, we use (*,G) state to describe bi-directional shared tree state (traditionally (*,G) has been used to describe uni-directional shared tree state).

By definition a PIM bi-directional shared tree group may not have any (S,G) state stored for the group. There are exceptions when mixing non-bidir PIM routers with bidir PIM routers (see later in this specification).

We assume that at the same time a router learns the RP for a group,

it will know if the group is to operate in bi-directional shared tree mode or uni-directional shared tree mode. This assumption greatly simplifies the deployment and operation of the protocol.

4. Modifications to Multicast Forwarding

There will be modifications to multicast forwarding since bi-directional shared tree delivery requires traffic to flow upstream (towards the root). This is contrary to RPF forwarding rules used on uni-directional shared trees (datagrams can only be forwarded away from the root node, downstream towards receivers).

5. No Modifications to Multicast Capable Hosts

This proposal does not require modifications to multicast capable hosts [3]. Hosts that receive multicast datagrams with the UMP option must ignore the option and accept the datagram [6].

6. How are hosts Joined to a Bi-directional Shared Tree

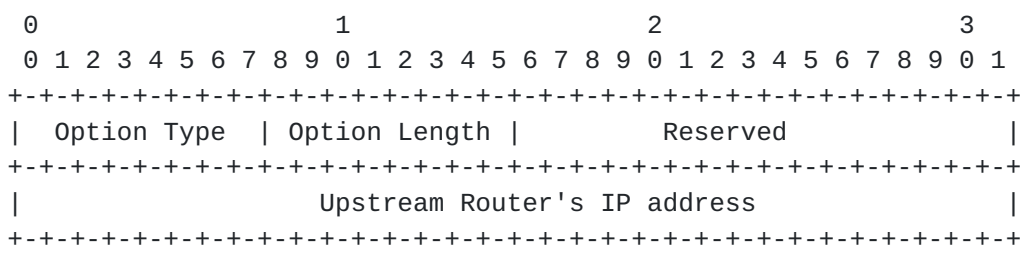
No change is required in hosts. Receiving hosts use IGMP [3] (as they do today) to join multicast groups. The attached designated router (DR) will initiate joining of the shared tree.

The attached routers perform the same actions as are done to graft branches on the uni-directional tree. That is, the designated router (DR), on the attached subnet with the receiver, will send a PIM Join/Prune message for (*,G) with the RP in the Join-List toward the RP for the group. Routers maintain (*,G) state as defined in the sparse-mode PIM specification. [1]

7. How are Source's Datagrams sent onto a Bi-directional Shared Tree

No change is required to hosts. A source host will send a multicast datagram by transmitting on its attached interface (as it does today) [3]. The attached DR will initiate the delivery of the multicast datagram upstream towards the RP.

When a datagram flows upstream, a receiving router must know that it can bypass the RPF check on the (*,G) entry. To accomplish this, we introduce a new IP option called the Upstream Multicast Packet (UMP). The UMP IP header option is encoded as follows:



Option Type value is 152 (0x98). That is, the 1-bit copied flag is set to 1, the 2-bit option class is set to 0, and the 5-bit option number is 24.

When a router forwards a multicast datagram upstream, it identifies the upstream router in the option. Only the indicated upstream router is responsible for forwarding the datagram upstream. When a router forwards a multicast datagram with the UMP option, it will multicast the datagram on the attached upstream subnet so other routers can forward datagrams down the shared tree if they have (*,G) state. Any directly attached members also receive the datagram.

In most cases, only a single copy of the datagram is sent upstream, taking advantage of multi-access/multicast capable media whereever possible. However, on the first hop LAN, there may be two datagrams that traverse the LAN. See next section for details.

It is important to note that symmetry between receivers and senders along the same branch must be maintained. That is, a router must join along the same path it would forward traffic upstream or loops could result. This specification forces symmetry because the same choice for forwarding and joining is achieved by using the RPF neighbor to the RP.

8. First-hop LAN

When a source transmits a multicast datagram, there is one router on the attached LAN that will insert the UMP option. The PIM designated router (DR) will be responsible for this. The DR will insert the UMP option using the address of the next-hop router it knows to reach the RP for the (*,G) entry.

There is one case where two datagrams traverse the first-hop LAN. The first datagram is transmitted as multicast by the source and the second is transmitted by the DR as MAC-level unicast to the next-hop upstream router. This only occurs when the DR uses the first-hop LAN as its RPF interface for (*,G). If the DR is an upstream router, the extra datagram is not sent because the RPF interface for (*,G) is not the first-hop LAN.

The DR is made responsible for selecting the upstream router in order to avoid inconsistent join and forwarding decisions if multiple downstream routers on the LAN receive joins or datagrams for the same group. If all routers on a LAN always ran a common link state protocol or always had some other means to guarantee consistent routing information, then this would not be necessary. However, in order to allow loop free operation in the widest range of environments, without making restrictive assumptions about unicast routing protocols, configurations and policies, we make use of the DR to enforce consistent decisions.

A network administrator can control which router is the PIM DR [5] to avoid particular suboptimal cases.

9. Multicast Forwarding Rules

The following steps describe the rules for bi-directional shared tree forwarding in PIM. When a router receives a multicast datagram, it may arrive on the RPF interface for a (*,G) entry or another (the non-RPF) interface.

A. When a multicast datagram arrives on the RPF interface toward the RP:

A1. A multicast routing table lookup is performed. Only a (*,G) entry can be returned (based on our definition that PIM bi-directional shared trees groups will not have (S,G) route entries).

If the entry is not found, the datagram is silently discarded.

A2. If the entry is found, the datagram is sent out each outgoing interface that resides in the outgoing interface list for the (*,G) entry. In this situation, the router doesn't care if the (*,G) tree state is bi-directional or uni-directional.

A3. Before replicating the datagram on each outgoing interface, a router checks to see if the UMP option is present. If so, it can either remove the option or replace the existing address with 0.0.0.0 in the Upstream Router IP Address field. This is to indicate to downstream routers the datagram is not traveling upstream.

A4. If the UMP option isn't present and the router is DR on the interface the datagram was received on and the source is directly attached, the DR is responsible for inserting the UMP option. It includes in the UMP option address the next-hop IP address of its RPF neighbor for (*,G). The DR forwards the datagram using the MAC-level address (unicast address) of this RPF neighbor.

A5. If the UMP option isn't present and the router isn't the DR, or the source isn't directly attached, the datagram is silently discarded.

B. When multicast datagram arrives on a non-RPF interface toward the RP:

B1. A multicast routing table lookup is performed. Only a (*,G) entry can be returned (based on our definition of PIM bi-directional shared trees).

B2. The router looks at the UMP option. If the option is present and the Upstream Router IP Address is not its own IP address on the received interface, the datagram is silently discarded.

B3. If the UMP option is not present and the router is directly connected to the source of the multicast datagram and is the DR on the interface, the DR inserts the UMP option and follows the steps B4.1 and B4.2.

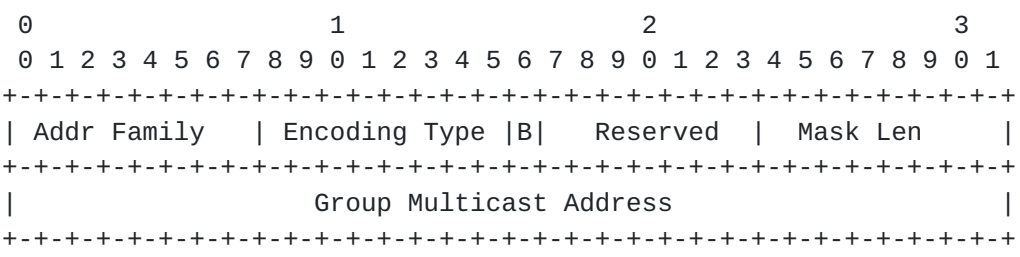
B4. If the UMP option is present and the Upstream Router IP Address field contains the IP address of the receiving router (on the received interface), it will forward the datagram as follows:

B4.1 The datagram is sent out each outgoing interface that resides in the outgoing interface list for the (*,G) entry except for the interface on which the datagram was received on. Before sending out each interface, the router may remove the UMP option from the datagram or replace the existing address with 0.0.0.0 in the Upstream Router IP Address field.

B4.2 The datagram is forwarded on the RPF interface for (*,G) by replacing the Upstream Router IP Address field in the UMP option with the next-hop address of the router that is used to reach the RP.

10. Distinguishing Bi-Directional Shared Tree Groups from other Groups

When routers discover the identity of the RP for a multicast group they can determine if the group will operate in bi-directional shared tree mode or uni-directional tree mode. We modify the Encoded-Group Address fields in PIM Bootstrap and Candidate-RP Advertisement messages to include the Bidir-bit (see bit B below):



When the Bidir-bit is set, all upgraded bi-directional PIM routers will follow the forwarding rules described in this specification.

11. Mixing Bi-Directional Capable with Uni-Directional-Only Routers

It will take time to upgrade all PIM routers in a domain to be bi-directional shared tree capable. However, enabling bi-directional shared tree routers in an existing network can be easy and simple. First, no special attention at the protocol level needs to be taken if the network is engineered where you can place bidir PIM routers strategically near sources. That is, if sources are located on sender-only branches (no Joins have traveled up that branch) of the bi-directional shared tree, only that branch needs to be upgraded with bi-directional shared tree capable routers. All other routers on receiver branches forward based on (*,G) uni-directional shared tree forwarding rules.

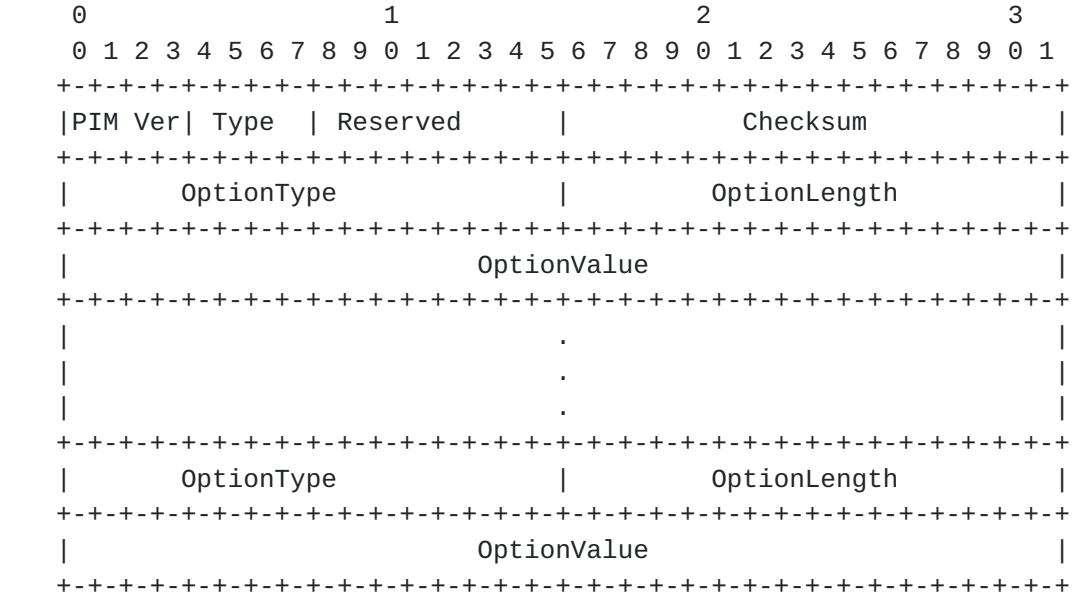
When the network cannot be engineered to locate bi-directional shared tree capable routers on sender-only branches, the following transition support can be implemented:

- A router will detect if its upstream neighboring router toward the RP is bi-directional shared tree capable or not. We will use the Bidir-Capable PIM Hello Option to convey this information.

- A router that is one-hop downstream (of the RP) from a non-bidir capable router will maintain (S,G) state and will be responsible for forwarding multicast traffic as to the RP by Registering to the RP as it would if it was a DR (or MBR) in uni-directional mode. When the data arrives at the RP, it can be delivered on the uni-directional shared-tree (or any source trees that overlap with the shared tree).

We define the following PIM Hello Option:

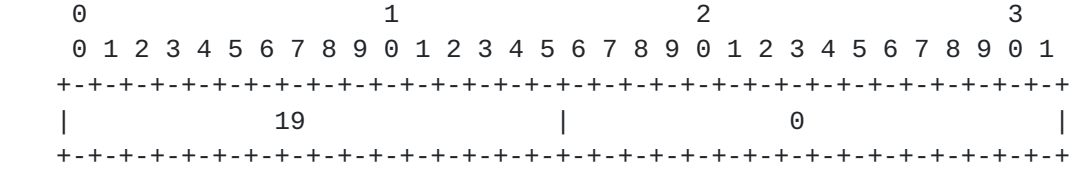
Bidir-Capable Option



OptionType = 19 (decimal)

OptionLength
The length of the option payload in bytes. This is set to 0.

Therefore, the following encoding is inserted in the PIM Hello message:



By definition, in a pure bi-directional router environment, a bidir capable PIM router will not create (S,G) state when it either 1) receives a datagram or 2) receives any PIM control message. However, there is one exception. When a router receives a datagram that is traveling upstream (the UMP option is present or the router is the DR directly attached to the source) and the upstream neighbor toward the RP is not bidir capable, it will create (S,G) state and set the necessary flags indicating datagrams that match the route entry will be Register encapsulated to the RP. In this case, the router still doesn't accept join messages (and therefore doesn't populate the (S,G) olist) if there are routers upstream that are sending (S,G) Joins or Prunes. A router that does this transition logic is called,

a bidir border router.

If a bidir router creates (S,G) state for a bi-directional group, it will not send Join/Prune messages for the entry. If a bidir router changes its RPF neighbor toward the RP and the RPF neighbor is bidir capable, it will delete its (S,G) entries.

A bidir router must do longest match lookups for a group that is in bi-directional tree mode. This handles the case where the RP forwards datagrams down a branch that has a both a sender and a member on it and avoids datagrams returning to the sender. In this case, a bidir border router should RPF fail for such datagrams since it will use the (S,G) entry rather than the (*,G) entry for the forwarding decision.

If the RP is a bidir capable router and it receives a Register message, it will not create (S,G) state. It will forward the data encapsulated in the Register message down the shared tree. The RP will only send a Register-Stop if there are no members for the group (the (*,G) outgoing interface list is empty). An RP will receive a Register message in two cases, 1) the DR is a non-bidir capable router, or 2) it was sent by a bidir border router.

If the RP is not bidir capable and it receives a Register from either a non-bidir capable DR or a bidir border router, it may trigger a Join toward the source. If there are any bidir capable routers on the path, they will not create (S,G) state. In this case, the RP will never get data natively and therefore never send Register-Stop messages. The data will continue to be delivered via Register encapsulation.

The following shows all possible cases mixing non-bidir (old) and bidir capable (new) routers. Each column shows the capability of 1) the RP, 2) an router between the DR and the RP, and 3) the DR. The following notation is used:

"new-rp" - RP is bidir capable
"old-rp" - RP is non-bidir capable
"old" - router is non-bidir capable
"new" - router is bidir capaple
"new-dr" - DR is bidir capable
"old-dr" - DR is non-bidir capable

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
old-rp	old-rp	old-rp	old-rp	new-rp	new-rp	new-rp	new-rp
old	old	new	new	new	new	old	old
old-dr	new-dr	old-dr	new-dr	new-dr	old-dr	new-dr	old-dr

(1) This case is uni-directional mode.

(2) The bidir DR sends Registers and does not forward datagrams with the UMP option. It does this because it detects the upstream router is not bidir-capable. The RP joins back to the source through the intermediate router. The intermediate router's join is ignored by the bidir DR. Datagrams get to receivers via Register encapsulation only from the DR.

(3) The non-bidir DR sends Registers. The non-bidir RP may send joins but the bidir intermediate router will ignore them. Datagrams get to receivers via Register encapsulation only from the DR.

(4) The bidir DR forwards multicast datagram with the UMP option upstream. It does this because it detects the upstream router is bidir-capable. The bidir intermediate router (acting as a bidir border router) sends Registers to the non-bidir RP. Datagrams get to receivers via Register encapsulation only from the bidir border router.

(5) This case is bi-directional shared tree mode.

(6) The non-bidir DR will Register to the bidir RP. The bidir RP will not send Joins back to the source. It only Register-Stops if there are no members. The bidir intermediate router is not involved in forwarding multicast datagrams. Datagrams get to receivers via Register encapsulation only from the DR.

(7) The bidir DR will Register to the bidir RP. It does this because it detects the upstream router is non-bidir capable. It is performing as a bidir border router. The bidir RP will not send Joins back to

the source. It only Register-Stops if there are no members. The non-bidir intermediate router is not involved in forwarding multicast datagrams. Datagrams get to receivers via Register encapsulation only from the DR.

(8) The non-bidir DR will Register to the bidir RP. The bidir RP will not send Joins back to the source. It only Register-Stops if there are no members. The non-bidir intermediate router is not involved in forwarding multicast datagrams. Datagrams get to receivers via Register encapsulation only from the DR.

12. Security Considerations

When IPsec [4] is used on a multicast datagram, the UMP IP option will not be part of the encrypted payload. This is justified by allowing routers to be performant when forwarding datagrams upstream.

13. Acknowledgments

The authors would like to acknowledge Rusty Eddy (USC), Radia Perlman (Sun), Tony Speakman (cisco), and Liming Wei (cisco) for their comments and contributions to this specification.

14. Author Information

Deborah Estrin
ISI/USC
estrin@usc.edu

Dino Farinacci
cisco Systems, Inc.
dino@cisco.com

15. References

[1] Estrin, et al., "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June 1998.

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

[3] Deering, S., "Host Extensions for IP Multicasting", STD 5, RFC 1112, August 1989.

[4] Atkinson, R., "Security Architecture for the Internet Protocol", RFC 1825, August 1995.

[5] Wei, L., Farinacci, D., "PIM Version 2 DR Election Priority Option", INTERNET-DRAFT, March 1998.

[6] ISI/USC, "Internet Protocol", RFC 791, September 1981.