Hierarchical PIM-SM Architecture for Inter-Domain Multicast Routing

Stephen Deering and Bill Fenner

Xerox PARC
3333 Coyoty Hill Road
Palo Alto, CA 94304
deering@parc.xerox.com,
fenner@parc.xerox.com

Deborah Estrin and Ahmed[1] Helmy

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

Dino Farinacci and Liming Wei

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

Mark Handley

Department of Computer Science
University College London
Gower Street
London, WC1E 6BT
UK
m.handley@cs.ucl.ac.uk

Van Jacobson

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

David Thaler

EECS Department
University of Michigan
Ann Arbor, MI 48109
thalerd@eecs.umich.edu

draft-ietf-idmr-Hierarchical-PIM.ps

December 18, 1995

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] Estrin and Helmy were supported by grants from the National Science Foundation and Sun Microsystems.

1 Introduction

In this document we present a hierarchical architecture for inter-domain multicast routing based upon PIM-SM protocol.

2 Goals

1. To attain the explicit join/receiver initiated behavior over the backbone. This is readily achieved by using PIM-SM.

2. To obtain Group to RP mapping information, at the PIM routers when needed, through algorithmic mapping, avoiding the delays and memory requirements for distributing and storing such information.

3. To retain scalability for the wide-area architecture. This requires that the number of entities stays manageable, without proportionally increasing with the size of the network; which leads evidently to a hierarchical structure.

   The proposed hierarchy deals with multi-level RPs; where for particular groups a higher level RP (level N) represents the point of rendezvous for all entities within a lower level (level N-1) in that portion of the hierarchy (see Figure 1) (where RP-level N is the Level N RP for multicast group, Gi, for the lower left branch of the hierarchy). Entities at level N-1 must have knowledge of level N RPs for their portion of the hierarchy.

   An open issue is how to form the hierarchy so that RPs know at what level of the hierarchy they reside and what portion of the hierarchy they serve. While initially we may rely on a configured hierarchy, a configured hierarchy is hard to maintain, and is not flexible enough to cope with the dynamics and the mechanistics involved. Thus, a self-configuring hierarchy would be better suited for the inter-domain case, in the longer term.

4. To utilize, to the largest extent possible, the mechanisms inherent in PIM-SM.

5. To utilize hierarchy to provide a layer of isolation between the different levels and help in achieving the modularity required.

3 Terminology

This section defines a number of terms used throughout this document. It may be skipped and used as a reference[2].

- **Administratively scoped groups**[3] are groups defined over a scope of an administrative region. Border routers enforce region constraint distribution for these groups. A specific range of the multicast address is allocated to such groups.

- **Bootstrap RPlist** is an RPlist used to bootstrap PIM-SM, by supporting essential well known groups used to advertise RP candidacy. A bootstrap RPlist is defined for each level of the hierarchy.

[2] Familiarity with basic PIM-SM terminology is assumed throughout this document.

[3] also known as 'scope limited', 'region-scoped' or simply 'local' groups, as opposed to 'global' groups.

- **Candidate-RP-Advertisement** (C-RP-Adv) is a PIM-SM message, sent periodically at low frequency by PIM-SM routers, which are configured to be candidate RPs. The C-RP-Adv message is level specific [i.e. contains information about candidacy for a certain level], and is unicast from the candidate RPs at one level to the active bootstrap RP at that same level.

  Advertisements for bootstrap candidacy are included in the PIM-Query messages; and hence are not distributed with the Candidate-RP-Advertisement messages.

- **Candidate-RP-Set** (C-RP-Set) message, is a message consisting of a set of Candidate-RPs, sent from the active bootstrap RP at level N, to the well-known Candidate-RP-Set-Advertisement group at that level.

- **Algorithmic mapping**, is an algorithm, possibly realized by a hash function, to obtain the RP information for a certain group.

- **Well Known Groups**

  The following Well Known Groups (WKGs) are needed:

  1. **Candidate-RP-set-Advertisement group(s):** A Candidate-RP-set-Advertisement (CaRPA) group is defined for each level of the hierarchy. The active bootstrap RP at each level sends Candidate-RP-Set messages to the corresponding CaRPA group.

     Entities at one level lower, (be they DRs at level 0, or RPs at one level lower), listen to the corresponding CaRPA group.

4 Basic Structure Overview

The basic structure consists of different levels of RPs[4]. Each level is labeled with a number (e.g. level N). RPs at level N-1 send joins/registers (as in standard PIM-SM) to the corresponding RP at level N; within one region, all RPs at the same level, within the same part of the hierarchy, converge (except for failures and transients) on a single higher level RP.

RPs at level N, on the other hand, send Reachabilities/Register-Acks to level N-1 (again, as in standard PIM-SM), (see figure 1).

5 Bootstrap Mechanism

The bootstrap mechanism proposed here, is the same as that used for intra-domain PIM-SM[1]. However, for inter-domain PIM-SM, a bootstrap RPlist is defined per level.

The bootstrap RPlist is used to deliver the Candidate-RP-Set messages for the corresponding level, within the corresponding region.

6 Isolating Multi-Level RP Interfaces

In order to decouple intra-domain protocol behavior from inter-domain developments as much as possible, we distinguish between the control messages exchanged at each level. This is realized by adding level flags in both the Reachability messages, and the Join/Prune messages. [Note that for Registers and Register-Acks the flags are not needed, as these messages are unicast.]

[4] DRs are considered as level 0 RPs.

Figure 1: Overview of the Hierarchy Architecture
Actions are numbered in the order they occur

1. **Reachability messages:**

   A level N RP processes a reachability message if the reachability message has level N flag set, indicating that the reachability message was sent from the level N+1 RP to level N RPs.

2. **Join/Prune messages:**

   Level N RPs send the Join/Prune messages towards the corresponding level N+1 RP, with the level N flag set. Level N Join/Prune messages build (*,G,level N) entries in the intermediate routers.

7 Resolving Overlaps

Note that when level I and level J (where I and J are different) trees overlap, a router may have more than a (*,G) entry with different settings for the incoming interface (iif) and outgoing interface list (oiflist).

These different (*,G) entries, however, will have different levels associated with them. To resolve this situation, the folowing rules are added:

1. Upon creating a (*,G) for a certain level (e.g. (*,G,level N) entry), a router copies the oiflist [excluding the iif(*,G,level N)] of entries (*,G,level M) [where M is less than N, if exist], into the oiflist of the newly created entry. Further, the router copies the oiflist(*,G,level N) into the oiflist of (*,G,level P) [where P is greater than N, if exist].

2. When a router has more than one (*,G) entry, it only uses the highest level one for data forwarding (see figure 2),

3. The router suppresses joins from being sent based on the lower level (*,G) entry(ies) (lower level means with an RP indicated that has a lower level value). However, it does not delete these lower level entry(ies)[5]. Instead, it periodically sends a null-register probing message to that level RP to

[5] note that however more than one (*,G) entry are present, only one forwarding cache is maintained in the kernel. All other entries, that trigger control messages, are kept at user level.

Figure 2: Resolving overlapping trees

detect its liveness. If the router receives a Register-ack from the RP, before the RP-timer expires, it refreshes the timer for that (*,G) entry. Otherwise, the RP-timer eventually times out.

[Note that the probing message refreshes the corresponding (*,G) entry at the RP, to maintain the joins to higher level RPs, even if there are no other downstream members at that level].

4. In addition, so long as the *,G (lower level) entry(ies) are alive, and so long as the router is getting RP liveness indication, the router generates RPreachability messages on behalf of the lower level RP(s) and forwards these down the *,G lower level tree(s). This allows DRs to continue to get lower level RP Reachabilities, even though they will get the data on the higher level part of the tree.

This avoids having to set up special forwarding state that would forward only reachabilities and not other data packets.

8 Allowing RP Overrides

In some worst case scenarios, where the shared tree is very suboptimal, it may be desirable to allow the group initiator to override the highest level RP for that group. This would allow for a better shared tree, and hence, better performance.

To achieve this, the following is performed:

1. The overriding RP sends a Candidate-RP-Advertisement to the active bootstrap RP at the highest level (i.e. LmaxN) (the highest level region should cover all regions over which that particular group is defined). The bootstrap RP then sends this C-RP-Adv with the C-RP-Set messages to the CaRPA group at the highest level. In so doing, it overrids the hash function for that particular group.

2. The next to highest level (or LmaxN-1) RPs adopt the overriding RP as their primary RP, and LmaxN RPs send RP-reachability messages indicating that DRs/RPs should use the new primary

RP[6].

This, however, incurs blackouts during the transition, and is thought of as an alternative to improve the shared tree for those applications having startup delay anyway.

9 Hierarchy

One main advantage of the hierarchy, is to provide locality. This is realized as follows[7]:

1. A Candidate RP at level N+1, sends periodic Candidate-RP-Advertisements to the active bootstrap RP at level N+1, which, in turn, sends periodic C-RP-Set messages, to nearby level N RPs, using corresponding CaRPA group, for that part of the hierarchy.

2. Nearby RPs at level N collect these C-RP-Set messages, and, according to a hash function (or similar) algorithm, map to the same level N+1 RP, for the corresponding group (or group prefix) address.

   The hash function acts upon the set of Candidate RPs, received from the bootstrap RP.

   An order-preserving hash function is called for, which allows to treat the Candidate RPs as a list of ordered RPs.

   Further, a mechanism may be introduced, through which, the hash function is updated in the pertinent PIM routers, if/when needed. This event is very infrequent, and is invoked only to achieve better load balancing and distribution of groups amongst candidate RPs.

10 Open Issues

10.1 Configuring the Hierarchy

As a first, short term, step, we propose to start with a statically configured hierarchy. In which, RPs at a certain level are configured with the RPs at one level higher.

The order-preserving hash function is introduced, thereafter, to map to the higher level RP(s).

An eventual goal would be to have a self configuring hierarchy, the RPs of which can, dynamically, distribute the candidacy information and establish the hierarchy at different levels, for the pertinent parts or regions.

References

[1] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, L. Wei, P. Sharma, and A. Helmy. Protocol independent multicast (pim): Specification. *Working Draft*, June 1995.

[6] This implies that level N RPs listen to CaRPA group for level N and level N+1, or simply, LmaxN RPs listen to CaRPA of LmaxN.

[7] We assume here that the hierarchy has already been configured and in place. How we achieve the hierarchy configuration is discussed in a following section