Network Working Group
Internet Draft
Expires: April, 1999

Dino Farinacci
Yakov Rekhter
cisco Systems
November 1998

Multicast Label Binding and Distribution using PIM
<draft-farinacci-multicast-tagsw-01.txt>

Status of this Memo

Abstract

This document describes a method for advertising labels for multicast flows. It strives to use downstream label assignment to be consistent with unicast label distribution. This proposal is media-type independent. Therefore, it works for multi-access/multicast capable LANs, point-to-point links, and NBMA networks.

1.0 Overview

We propose to use PIM and combine the (*,G) and (S,G) join state with label assignment and distribution. Labels and multicast routes will be sent together in one message.

1.1 Goals

i. We are motivated to have the upstream Label Switch Router (LSR)

use one label for multicast data delivery on a network so we can make use of data-link multicast delivery where available.

ii. We are motivated to use downstream label assignment to achieve:

- o Simplicity and consistency with unicast label assignment.
- o A per interface Label Information Base (LIB) that guarantees unique label assignments on any interface.
- o Consistent algorithms for label assignment and distribution among different media types.
- o Both routing table state and the label binding information associated with the state are advertised together in a single control message thus reducing race conditions.
- o Avoid label reallocation or reassignment when there are RPF changes (i.e. the multicast distribution tree takes different shape).
- o To improve utilization of label space by randomizing label assignment among all downstream routers joining for a group.

iii. Works with dense-mode or sparse-mode operation.

2.0 Proposal

A LSR that supports multicast sends PIM Join messages on behalf of hosts that join groups. It sends Joins messages to upstream neighboring LSRs toward the RP for the shared-tree (*,G) or toward a source for a source-tree (S,G). If the LSR creates the state for the group, it will assign a label for the respective (*,G) or (S,G) state. It includes the label in the Join message associated with the multicast routing table entry. The entry is created in its LIB using the label as its incoming label component.

The upstream LSR, when it receives the Join, will cache the new multicast routing table state along with the label. An entry is created in the LIB and the label is used as the outgoing component. This label will be used by the upstream LSR to forward multicast data packets.

Since PIM Join messages are multicast on a LAN, other downstream LSRs, that are interested in the group, will hear the message and can cache the binding of multicast routing table state and label state together. Since the upstream LSR is going to forward data packets

using the advertised label, they must be ready to accept the data packet with that advertised label.

The first downstream LSR that joins for a group, is the label assigner (or called in other forums as the Label Allocation Server) on a LAN for a multicast route. All other downstream LSRs that send PIM Join messages will use the same label that the assigner selected. A LSR that sends a PIM Join message with a label of 0 means that it doesn't know the label for the associated multicast routing table entry. When this occurs, the assigner can trigger a PIM Join message making the label known.

This algorithm works on point-to-point links because there is only one downstream LSR on the link which always becomes the label assigner.

On NBMA networks, all PIM routers are known to each other through pseudo-broadcast mechanisms provided by the data-link layer. However, PIM Join messages are unicast to the upstream LSR. Therefore, other downstream LSRs will not hear the label assigner's advertisement. To overcome this issue, we have each downstream LSR become the label assigner on NBMA networks. Since the upstream LSR is going to pseudo-broadcast the data anyways it can supply a label for each packet that goes to each respective downstream LSR.

2.1 Corner cases

Multiple downstream LSRs cannot assign the same label value for any multicast route because they partition the label space into non-overlapping ranges according to [4]. When a LSR is enabled on an interface, it obtains a unique label range for the LAN.

When the label assigner leaves the group, the label that it assigned still remains active. The next highest IP addressed downstream LSR becomes the owner of that label and may change it if it sees fit.

However, it is not required to change it. All downstream LSRs can continue to use the assignment in their Join messages.

If two systems both join for the first time (they do not have state), at the same time and each choose a different label value, the highest IP addressed downstream LSR's label will be used by the upstream LSR. The lower addressed LSR will hear the higher addressed LSR's Join too and will also use it's label.

If the label assigner crashes, the highest IP addressed downstream LSR assigns a new label to the multicast routes, which were assigned by the crashing LSR, and triggers a Join message so all other LSRs on

the LAN to use the new label.

When a LAN partitions due to a layer-2 switch failure, it follows the same logic for the case when a LSR stops joining for a group. When the partition heals, there may be an RPF neighbor change in one of the partitions. When there is an RPF neighbor change and the downstream routers trigger joins to their new RPF neighbor with a different label assignment than the other partition is using, one of two resolutions occur:

1) The LSR which is the allocator in the partition of the new RPF neighbor will trigger a join if it has a higher IP address than the allocator in the other region. The downstream routers in the other partition use the new label assignment immediately.

2) If the LSR which is the allocator in the partition of the new RPF neighbor has a lower IP address, all downstream routers and the new RPF neighbor will switch to the label assigned by the allocator in the other partition.

If an RPF change occurs (the topology changed so the upstream LSR is different), the PIM protocol spec indicates that a PIM Join may be triggered to get on the new distribution tree as soon as possible. In this case, if the label assigner becomes the upstream LSR, then the new highest IP addressed downstream LSR may become the label assigner. It may change the label if it sees fit. Otherwise, the same label is used.

3.0 Coexistence of Label-Capable and Label-Incapable multicast routers

An upstream router will know if all routers on a subnet are LSRs or not. If there are any label incapable routers, the upstream router will not label encapsulate multicast data packets. The PIM Hello message will indicate if the router is label capable. The PIM Hello message is sent by every multicast capable router.

If the upstream router detects any non-PIM neighbors on the subnet, it will assume that they are label capable and will not label encapsulate multicast data packets.

An optimization may be achieved, if the upstream router knows that all downstream routers interested in the group are LSRs, it may label encapsulate multicast data packets even though there are other label incapable routers on the subnet.

Related to the above cases, if there is a group member on a LAN, co-located with a multicast LSR, only a single packet will be forwarded.

It is the responsibility of the upstream router to decapsulate the labeled packet and forward it on the LAN as an IP packet so the member can receive it. The downstream routers may forward the IP packet or label encapsulate it.

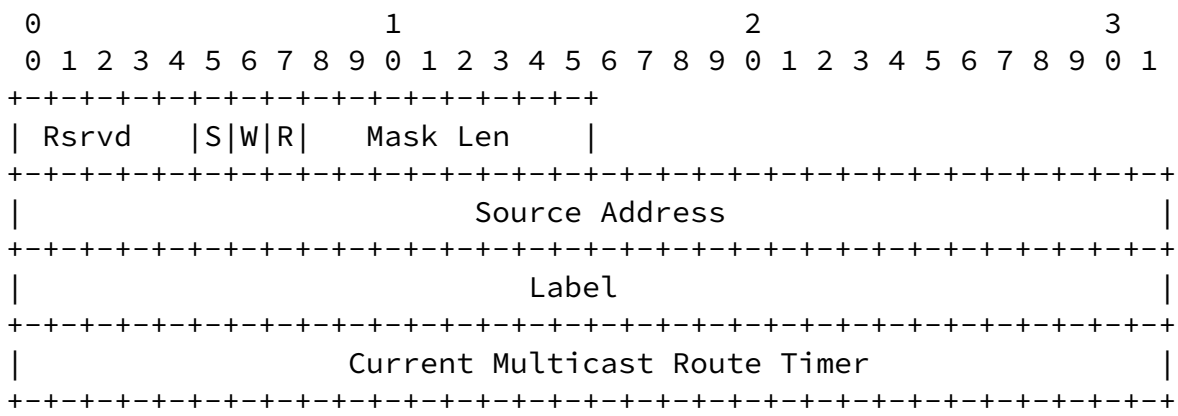
4.0 Label Conflict Resolution

The use of different data-link layer code-points (i.e. Ethertypes, PPP protocol types) for unicast and multicast label switching allows to disambiguate between labels associated with unicast routes versus labels associated with multicast routes. Therefore, the assignment of labels for unicast routes could be done completely independent from the assignment of labels for multicast routes, without creating any risk of ambiguity. For example, the same label value could be allocated for a unicast route and for a multicast route.

5.0 Modifications to PIMv2

PIMv2 has a packet format for each address type it may support when encoding both multicast and unicast addresses. We will define a new

address type called "Label Address" for unicast address encoding. The label will accompany the source address in the Encoded Source Address format as specified in [2]. The label value will be in a 32-bit quantity following the source address. So, for example, an IPv4 Label Address format would look like:



Label
A 20-bit value assigned by the LSR sending the Join/Prune message. The 20 bit value is inserted in the low-order 20-bits of this 32-bit field.

Current Multicast Route Timer
The sender of a Join/Prune message inserts the current time left before expiration for the multicast route table entry described by

the Source Address (either the (S,G) or (*,G) entry). This is needed so all routers on a common multi-access subnet can time-out the entry close to the same time without each other recreating the state when the source goes inactive.

Refer to [2] for other field descriptions not specified here.

6.0 Label Distribution for dense-mode groups

In dense-mode PIM, there is no downstream Join message traveling upstream to perform the binding of multicast routes with labels. However, since we don't want a separate algorithm for dense-mode groups, we extend this basic design for dense-mode PIM.

When a downstream LSR creates (S,G) state from the receipt of 1) data, or 2) Join/Prune or Graft messages, it will start a periodic timer to send Join messages with label assignment information present. The messages look no different and are treated on receipt no differently than in the sparse-mode case.

The periodic Join message will be multicast on the LAN with an upstream target address of 0.0.0.0. All multicast LSRs on the LAN must know the group operates in dense-mode. This is accomplished using standard PIM mechanisms.

7.0 Security Considerations

Security considerations are not discussed in this memo.

8.0 Acknowledgments

The authors would like to thank Fred Baker and Eric Rosen from cisco Systems for their insightful comments on this draft.

9.0 Author's Address

Dino Farinacci
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134
Email: dino@cisco.com

Yakov Rekhter
Cisco Systems, Inc.

170 Tasman Drive
San Jose, CA, 95134
Email: yakov@cisco.com

10.0 References

[1] Multiprotocol Label Switching Architecture, draft-ietf-mpls-

arch-02.txt, Rosen, Viswanathan, Callon, July, 1998.

[2] Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification, RFC 2362, Estrin, Farinacci, Helmy, Thaler, Deering, Handley, Jacobson, Liu, Sharma, Wei, June, 1998

[3] LDP Specification, <draft-ietf-mpls-ldp-01.txt>, Andersson, Doolan, Feldman, Fredette, Thomas, August, 1998

[4] Partitioning Label Space amoung Multicast Routers on a Common Subnet, <draft-farinacci-multicast-tag-part-01.txt>, Farinacci, October, 1998

[5] "MPLS Label Stack Encoding", draft-ietf-mpls-label-encaps-03.txt, Rosen, Rekhter, Tappan, Fedorkow, Li, Conta, September, 1998