Network Working Group
Internet Draft
Expires: June, 1997

Dino Farinacci
Yakov Rekhter
cisco Systems
December 1996

Multicast Tag Binding and Distribution using PIM
<draft-farinacci-multicast-tagsw-00.txt>

Status of this Memo

Abstract

This document describes a method for advertising tags for multicast flows. It strives to use downstream tag assignment to be consistent with unicast tag 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 tag assignment and distribution. Tags and multicast routes will be sent together in one message.

1.1 Goals

i. We are motivated to have the upstream Tag Switch Router (TSR) use

one tag 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 tag assignment to achieve:

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

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

2.0 Proposal

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

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

Since PIM Join messages are multicast on a LAN, other downstream TSRs, that are interested in the group, will hear the message and can cache the binding of multicast routing table state and tag state together. Since the upstream TSR is going to forward data packets using the advertised tag, they must be ready to accept the data

packet with that advertised tag.

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

This algorithm works on point-to-point links because there is only one downstream TSR on the link which always becomes the tag 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 TSR. Therefore, other downstream TSRs will not hear the tag assigner's advertisement. To overcome this issue, we have each downstream TSR become the tag assigner on NBMA networks. Since the upstream TSR is going to pseudo-broadcast the data anyways it can supply a tag for each packet that goes to each respective downstream TSR.

2.1 Corner cases

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

When the tag assigner leaves the group, the tag that it assigned still remains active. The next highest IP addressed downstream TSR becomes the owner of that tag and may change it if it sees fit. However, it is not required to change it. All downstream TSRs 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 tag value, the highest IP addressed downstream TSR's tag will be used by the upstream TSR. The lower addressed TSR will hear the higher addressed TSR's Join too and will also use it's tag.

If the tag assigner crashes, the highest IP addressed downstream TSR assigns a new tag to the multicast routes, which were assigned by the crashing TSR, and triggers a Join message so all other TSRs on the LAN to use the new tag.

When a LAN partitions due to a layer-2 switch failure, it follows the same logic for the case when a TSR 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 tag assignment than the other partition is using, one of two resolutions occur:

1) The TSR 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 tag assignment immediately.

2) If the TSR 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 tag assigned by the allocator in the other partition.

If an RPF change occurs (the topology changed so the upstream TSR 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 tag assigner becomes the upstream TSR, then the new highest IP addressed downstream TSR may become the tag assigner. It may change the tag if it sees fit. Otherwise, the same tag is used.

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

An upstream router will know if all routers on a subnet are TSRs or not. If there are any tag incapable routers, the upstream router will not tag encapsulate multicast data packets. The PIM Hello message will indicate if the router is tag 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 tag incapable and will not tag encapsulate multicast data packets.

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

Related to the above cases, if there is a group member on a LAN, co-located with a multicast TSR, only a single packet will be forwarded. It is the responsibility of the upstream router to decapsulate the tagged 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

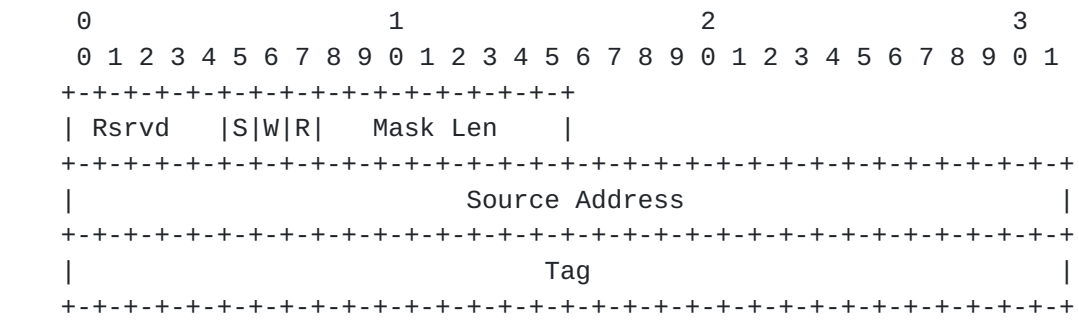
tag encapsulate it.

4.0 Tag Conflict Resolution

The use of different data-link layer code-points (i.e. Ethertypes, PPP protocol types) for unicast and multicast tagswitching allows to disambiguate between tags associated with unicast routes versus tags associated with multicast routes. Therefore, the assignment of tags for unicast routes could be done completely independent from the assignment of tags for multicast routes, without creating any risk of ambiguity. For example, the same tag 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 "Tag Address" for unicast address encoding. The tag will accompany the source address in the Encoded Source Address format as specified in [2]. The tag value will be in a 32-bit quantity following the source address. So, for example, an IPv4 Tag Address format would look like:



Refer to [2] for field descriptions.

6.0 Tag 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 tags. 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 TSR 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 tag 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 TSRs 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] Tag Switching Architecture Overview, draft-rfced-tag-switching-overview-00.txt, Rekhter, Davie, Katz, Rosen, Swallow

[2] Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification, <draft-ietf-idmr-pim-sm-spec-09.txt>, Estrin, Farinacci, Helmy, Thaler, Deering, Handley, Jacobson, Liu, Sharma, Wei, October, 1996

[3] Tag Distribution Protocol, <draft-doolan-tdp-spec-00.txt>, Doolan, Davie, Katz, Rekhter, Rosen, September, 1996

[4] Partitioning Tag Space amoung Multicast Routers on a Common Subnet, Farinacci, December, 1996

[5] "Tag Switching: Tag Stack Encodings", <draft-rosen-tag-stack-00.txt>, Rosen, Tappan, Farinacci, Rekhter, Fedorkow, November, 1996