pim
Internet-Draft
Updates: 3810, 4541 (if approved)
Intended status: Standards Track
Expires: 16 September 2025

N. Karstens
Garmin International
Z. Zhang
L. Giuliano
N. Ashik
Juniper Networks
J. Huang
Garmin International

15 March 2025

Multicast Snooping Optimizations
draft-karstens-pim-multicast-snooping-optimization-01

Abstract

TODO: provide abstract

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 16 September 2025.

Copyright Notice

Table of Contents

1. Introduction

Considerations for the operation of IGMP and MLD snooping switches are described in [SNOOP]. In the intervening years since publication, industry has gained experience with deploying this technology and have identified areas of improvement on the original document, including how traffic distribution can be optimized for certain types of networks.

One area of improvement is that there appears to be a gap in the control path forwarding rules outlined in [SNOOP], Section 2.1.1. Forwarding rules are defined for switch ports attached to multicast routers and switch ports attached to hosts, but switch ports attached to other switches are not mentioned, leaving the operation of these ports open to interpretation.

The authors may have purposefully limited consideration to networks with only a single switch, though this is not explicitly stated. In one sense, a network with a hierarchy of switches may be modeled as a

single switch (in other words, the fact that multiple switches are being used would be transparent to any host or multicast router attached to the network). One side-effect of this approach is that it obscures how the network distributes traffic between multiple switches.

This router-centric view of multicast traffic distribution is likely rooted in the nature of the IGMP and MLD protocols, whose stated purpose is to communicate group membership (that is, the fact that a host would like to receive a given stream of multicast data) to a multicast router. Two types of networks would benefit from a closer examination of how traffic is distributed between multiple switches: 1) networks that do not contain a multicast router and 2) networks that contain a multicast router, but have a significant volume of multicast data that does not need to be routed outside of the local network.

The following diagram depicts an example network that can be used to illustrate the point:

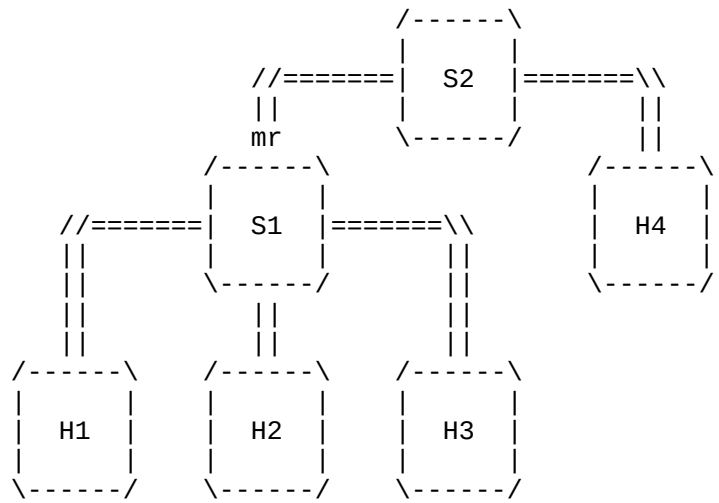


Figure 1: Example Network

S1 has designated its port connected to S2 as a multicast router port.

Suppose the example network contains two multicast streams:

* Stream 1 generated by H1 and consumed by H3

* Stream 2 generated by H2 and consumed by H4

The problem is that Stream 1 is forwarded from S1 to S2 even though there are no consumers of this data. This is due to the following data forwarding rule in [SNOOP], Section 2.1.2:

> Packets with a destination IP address outside 224.0.0.X which are not IGMP should be forwarded according to group-based port membership tables and must also be forwarded on router ports.

While it would be tempting to ignore this rule so that Stream 1 is no longer forwarded, this would also prevent Stream 2 from reaching H4. This is because of the following IGMP forwarding rule in [SNOOP], Section 2.1.1:

> A snooping switch should forward IGMP Membership Reports only to those ports where multicast routers are attached.

This rule prevents S2 from forwarding the Membership Report from H4 to S1, so S1 does not mark the port connected to S2 as a member of the multicast group for Stream 2.

[SNOOP] indicates that this rule is meant to prevent a host running IGMPv1 or IGMPv2 (or MLDv1) from suppressing its Membership Report for that multicast group. While it would be tempting to require all hosts on the network to run IGMPv3 (or MLDv2), this is not possible on the networks targeted for deployment of this solution.

The next logical step in this direction would be to require multicast snooping switches to use some method to identify the nature of the device connected to each port (switch or host, IGMP/MLD version, etc.). This information would then be used to help control the distribution of Membership Reports.

However, this document uses a different approach, choosing instead to use General Query messages to ensure membership information is distributed to all multicast snooping switches on the network. This solution is less complicated than methods that focus on Membership Reports, which reduces the likelihood for error. In addition, compatibility between different versions of IGMP and MLD are less of a concern because each protocol’s Query messages are compatible with earlier versions of the protocol.

Multicast snooping switches implementing this design shall use [IGMPv3], [MLDv2], or both. Multicast routers on the network must also use these protocol versions, [PIM-SM], and satisfy the requirements listed in Section 4.2. Network hosts may use earlier versions of the protocols.

1.1. Terminology

This document refers to IGMP snooping and MLD snooping collectively as "multicast snooping".

TODO: if there are notable differences then mention something like "except where there are notable differences".

IGMPv3 Membership Report messages and MLDv2 Multicast Listener Report messages are referred to collectively as Reports.

The terms Any-Source Multicast and Source-Specific Multicast (see [SSM]) are respectively abbreviated ASM and SSM.

TODO: do we want to define "multicast snooping switch"? Should we have an abbreviation?

2. Proxy Query Messages

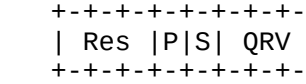
[SNOOP], Section 2.1.1 describes a special-case IGMP Query message with an IPv4 source address of 0.0.0.0. It would appear that the equivalent for MLD would be a Query message with the IPv6 source address set to the unspecified address (::), but several documents prohibit this:

* [RFC2710], Section 3 requires all MLD messages to be sent with a IPv6 link-local source address

* [RFC3590], Section 4 requires MLD Query messages to be sent with a valid IPv6 link-local source address

* [MLDv2], Section 5.1.14 not only requires MLD Query messages to be sent with a valid IPv6 link-local source address, but also that nodes MUST discard Query messages with an IPv6 source address that is not a valid IPv6 link-local address

Instead of working against established precedent, this document modifies the Query message described in [MLDv2], Section 5.1, repurposing the reserved bit immediately preceding the S flag as a new P (Proxy Query) flag:



If an IPv6 multicast router receives a Query with the P flag set, then it SHALL NOT use that message in the querier election process described in [MLDv2], Section 7.6.2.

Note that this document does not reassign the corresponding bit in the IGMP Query message ([IGMPv3], Section 4.1). That message only has four reserved bits, so it seemed better to leave that bit available for future use.

This document uses the term Proxy Query to refer to an IGMP Query with an IPv4 source address of 0.0.0.0 or an MLD Query with the P flag set.

3. Control Plane Operations

Multicast snooping switches shall maintain a group-based port membership table that indicates which port(s) contain members for each tracked group. The requirements in this section are ultimately related to managing this table, using IGMP/MLD to communicate its contents to adjacent nodes, and making changes in response to IGMP/MLD messages received from adjacent nodes.

Multicast snooping switches shall maintain a new Group/Port Membership Interval timer for each group and port combination in the group-based port membership table (see Section 5.1).

TODO: describe when the timer is set/reset and what happens when it expires

Multicast snooping switches shall track a Multicast Router flag for each port. This flag indicates that a multicast router is connected through the associated port.

Multicast snooping switches shall also track a Proxy Querier flag for each port. This flag indicates that the switch has received a Proxy Query message from this port, likely from another multicast snooping switch.

If neither the Multicast Router or Proxy Querier flag is set for a port, then it can be inferred that the port is connected either to a host or to a switch that does not implement this design.

3.1. Transmitting Periodic Proxy Queries

Each multicast snooping switch on the network shall periodically send General Proxy Query messages to all active ports. On startup, or after a port transitions from an inactive state to an active state, [Startup Switch Proxy Query Count] (see Section 5.4) General Proxy Query messages shall be sent with [Startup Switch Proxy Query Interval] (see Section 5.3) between each message. After this, General Proxy Query messages shall be sent with [Switch Proxy Query Interval] (see Section 5.2) between each message.

The QQIC field for each General Proxy Query message shall be set to [Startup Switch Proxy Query Interval] or [Switch Proxy Query Interval], as appropriate (see [IGMPv3], Section 4.1.7 or [MLDv2], Section 5.1.9).

TODO: Important to send to multicast router ports because multicast router may have IRB (Integrated Routing and Bridging) connected, or be running a PIM-to-IGMP proxy.

3.2. Receiving Queries

Multicast snooping switches shall do the following when a General non-Proxy Query is received:

1. Set the Multicast Router flag for the port

2. Reset the Multicast Router Timeout timer for the port (see Section 5.5)

3. Forward the Query to all other active ports

Multicast snooping switches shall do the following when a General Proxy Query is received:

1. Set the Proxy Querier flag for the port

2. Reset the Proxy Querier Timeout timer for the port (see Section 5.6)

Note that Proxy Query messages are not forwarded. Forwarding non-Proxy Queries facilitates querier election and ensures that the network is aware that a multicast router is present. Accordingly, if only Proxy Query messages are received, then it can be inferred that there is no multicast router on the network.

Multicast snooping switches shall respond to either type of Query by sending a Report to the port the Query was received on. This Report contains all groups in the group-based port membership table except the groups where the port the Query was received on is the only member port.

TODO: we can add a table that gives an example group-based port membership table and the contents of the resulting report

TODO: Does it make sense to require General Query messages be forwarded to multicast router ports, while also implying that if only Proxy Queries are received that there is no multicast router?

TODO: use a term like "Querier Ports" to indicate a port that a Proxy Query is received on. Update: look for Proxy Querier flag

TODO: need to handle unsolicited Rrports. That needs to be sent to Querier Ports

TODO: need to handle leave and group-specific queries -- snooping switches will keep track of the ports that have group membership. When it receives a leave then it subtracts the port from membership, but only forwards the leave if there are no other ports with group membership

TODO: group membership timer expiration is the same as an explicit leave

TODO: if a switch receives a leave and there is only one port remaining and that port is a querier port, then send a leave to that querier port

3.3. Receiving Reports

TODO

3.4. Transmitting Reports

TODO

3.5. Removing Groups from the Membership Table

TODO

4. Data Plane Operations

Multicast snooping switches shall follow the data forwarding rules outlined in [SNOOP], Section 2.1.2. In order to optimize traffic distribution on the network, this section contains two refinements to the recommendation to forward traffic to the multicast router port, based on whether the multicast traffic is routable or non-routable.

To some extent, what constitutes a routable multicast address is subject to the overall design of the network, so multicast snooping switch vendors should allow configuration for how multicast addresses are classified.

Note that the decision to forward traffic based on group-based port membership tables is independent of the decision to forward traffic to the multicast router port. In other words, traffic may still be forwarded to the multicast router port because it is a group member.

4.1. Non-Routable Multicast Traffic

Multicast snooping switches shall only forward traffic to the multicast router port if the traffic is known to be routable.

[SNOOP], Section 3 discusses challenges associated with IPv6 addresses overlapping when they are mapped to DMAC addresses. Multicast snooping switches should account for this possibility when implementing this requirement. In the event of an address collision, the recommendation is to forward the traffic and alert the network administrator of the problem.

TODO: How should this alert work, is there a YANG model we should update?

4.2. Routable Multicast Traffic

Multicast snooping switches shall begin in a state where all routable, ASM traffic is sent to the multicast router, which will route it outside of the network. When the traffic reaches the RP, the RP determines if it is interested in the traffic. If the RP is not interested in the traffic, then it will send a PIM prune message back to the multicast router.

When the multicast router receives the PIM prune message, it shall send a Report message excluding the multicast address back to the multicast snooping switch. Multicast snooping switches shall propagate the exclusion message back toward the source of the multicast stream until it reaches a switch that contains a port that is a member of that multicast group.

The fact that an exclusion message was received on the multicast router port should be recorded so that the network can be properly updated in the case of future changes to group-based port membership. For example, if a multicast snooping switch stops propagating an exclusion message because it contains at least one port that is a member of the multicast group, then the switch should send an exclude message back towards the source of the multicast stream when the last port is removed from membership in the multicast group.

TODO: is this a problem? we woule be preventing future refinements where an exclude message could be used to leave the have the port leave group membership

Note that SSM traffic is handled differently because PIM will send a request to receive the traffic, so the recommendations in this section only apply to ASM traffic.

5. List of Timers, Counters and Their Default Values

TODO: Add a note indicating that these should be consistent with the analagous timers on a single link (like IGMPv3 section 8 or MLDv2 section 9 requires)

5.1. Group/Port Membership Interval

This interval is analagous to the Group Membership Interval in [IGMPv3], Section 8.4 and Multicast Address Listening Interval in [MLDv2], Section 9.4, except it denotes how long a given group/port combination should remain in the group-based port membership table.

5.2. Switch Proxy Query Interval

This interval is analogous to the Query Interval in [IGMPv3], Section 8.2 and [MLDv2], Section 9.2, except it denotes the interval between General Proxy Queries sent by the multicast snooping switch.

5.3. Startup Switch Proxy Query Interval

This interval is analogous to the Startup Query Interval in [IGMPv3], Section 8.6 and [MLDv2], Section 9.6, except it denotes the interval between General Proxy Queries sent by the multicast snooping switch at startup.

5.4. Startup Switch Proxy Query Count

This interval is analogous to the Startup Query Count in [IGMPv3], Section 8.7 and [MLDv2], Section 9.7, except it denotes the number of General Proxy Queries sent on startup, separated by the Startup Switch Proxy Query Interval.

5.5. Multicast Router Timeout

This timeout is somewhat analogous to the Other Querier Present Timeout timer in [IGMPv3], Section 8.5 and [MLDv2], Section 9.5. If the timer expires, then the Multicast Router flag is cleared for the associated port.

5.6. Proxy Querier Timeout

This timeout is somewhat analogous to the Other Querier Present Timeout timer in [IGMPv3], Section 8.5 and [MLDv2], Section 9.5. If the timer expires, then the Proxy Querier flag is cleared for the associated port.

6. Security Considerations

To be added.

7. IANA Considerations

This document does not have any IANA assignments/requests.

8. References

8.1. Normative References

[IGMPv3] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <https://www.rfc-editor.org/info/rfc3376>.

[MLDv2] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <https://www.rfc-editor.org/info/rfc3810>.

[PIM-SM] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

[SNOOP] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, DOI 10.17487/RFC4541, May 2006, <https://www.rfc-editor.org/info/rfc4541>.

[SSM] Bhattacharyya, S., Ed., "An Overview of Source-Specific Multicast (SSM)", RFC 3569, DOI 10.17487/RFC3569, July 2003, <https://www.rfc-editor.org/info/rfc3569>.

8.2. Informative References

[RFC2710] Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, DOI 10.17487/RFC2710, October 1999, <https://www.rfc-editor.org/info/rfc2710>.

[RFC3590] Haberman, B., "Source Address Selection for the Multicast Listener Discovery (MLD) Protocol", RFC 3590, DOI 10.17487/RFC3590, September 2003, <https://www.rfc-editor.org/info/rfc3590>.

Acknowledgements

The authors would like to recognize the following individuals for their contributions to this research:

* David Vandewalle
  Garmin International

* Princy Elizabeth
  Juniper Networks

Authors' Addresses

Nate Karstens
Garmin International
Email: nate.karstens@garmin.com

Zhaohui Zhang
Juniper Networks
Email: zzhang@juniper.net

Lenny Giuliano
Juniper Networks
Email: lenny@juniper.net

Naveen Ashik
Juniper Networks
Email: nashik@juniper.net

Joseph Huang
Garmin International
Email: joseph.huang@garmin.com