Internet Engineering Task Force
INTERNET-DRAFT
draft-ietf-idmr-membership-reports-01.ps

Inter-Domain Multicast Routing WG
W. Fenner
August 5, 1998
Expires December 1998

Domain Wide Multicast Group Membership Reports

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."

To learn the current status of any Internet-Draft, please check the "1id-abstracts.txt" listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa), nic.nordu.net (Europe), munnari.oz.au (Pacific Rim), ds.internic.net (US East Coast), or ftp.isi.edu (US West Coast).

Distribution of this document is unlimited.

Abstract

When running a multi-level multicast routing protocol, upper levels need to know about group memberships in lower levels in a protocol-independent fashion. Domain Wide Multicast Group Membership Reports allow this information to be learned in a fashion similar to IGMP[Fenn97] at the domain level.

This document is a product of the IDMR working group within the Internet Engineering Task Force. Comments are solicited and should be addressed to the working group's mailing list at idmr@cs.ucl.ac.uk and/or the author.

1. Introduction

Domain-Wide Multicast Group Membership Reports (DWRs) are a group membership protocol at the domain level. When using a hierarchical multicast routing protocol [Thya94,Estr98], the inter-domain protocol needs to learn of group memberships inside domains. Although some intra-domain routing protocols can provide this information easily to the domain border routers, some cannot. DWRs specify a protocol-independent manner to learn group membership inside a domain.

This document specifies the DWR protocol, as used by border routers and interior routers. It specifies a behavior that can be used with any intra-domain protocol, along with optimizations for certain intra-domain protocols, and a transition scheme so that all interior routers need not be updated.

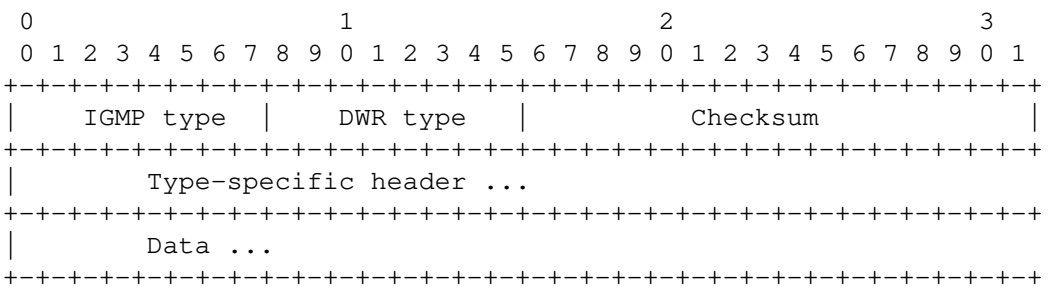
1.1. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [Bradner97].

Tunable timer values are named inside square brackets, e.g. [Robustness Variable]. These values are described in section 7.

2. Packet Format

DWR packets are sent as IGMP packets (IP protocol #2) when using IPv4. It's not yet clear what IPv6 type will be used, or if they should have their own IP protocol number which can be the same for v4 and v6.



2.1. IGMP Type: 8 bits

The IGMP type field is defined to be 0x21.

2.2. DWR Type: 8 bits

The following DWR types are defined:

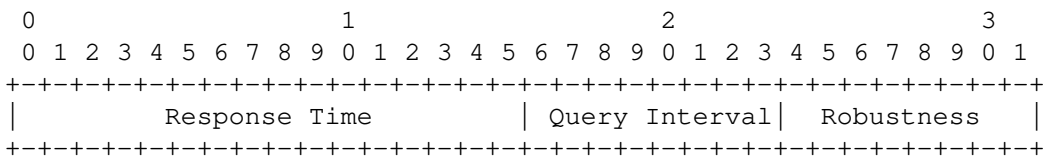
Type	Name
0x00	Domain-Wide Query
0x01	Domain-Wide Membership Report
0x02	Domain-Wide Leave
0x03	Non-authoritative Domain-Wide Leave

2.3. Checksum

IP-style checksum [Brad88] over the whole packet, zeroed for computing checksum. This checksum MUST be computed when transmitting packets, and MUST be verified when receiving. Received packets with incorrect checksums MUST be dropped.

2.4. Type-specific header

The only type-specific header defined is for the Domain-Wide Query; its header contains:



2.4.1. Response Time

The time, in units of 10ms, allowed for responses to this query. Varying this field allows tuning the burstiness of DWR traffic at the cost of higher latencies.

2.4.2. Query Interval

The time, in units of 10 seconds, between periodic Query messages from this Querier.

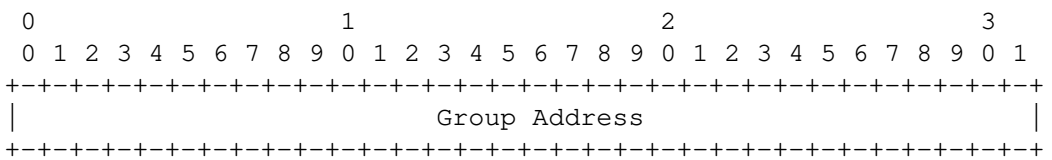
2.4.3. Robustness

The Robustness variable, described later. Along with the Query Interval, conveying this data in the Query allows exact calculation of Querier timeouts and allows interior routers to calculate the group membership lifetime.

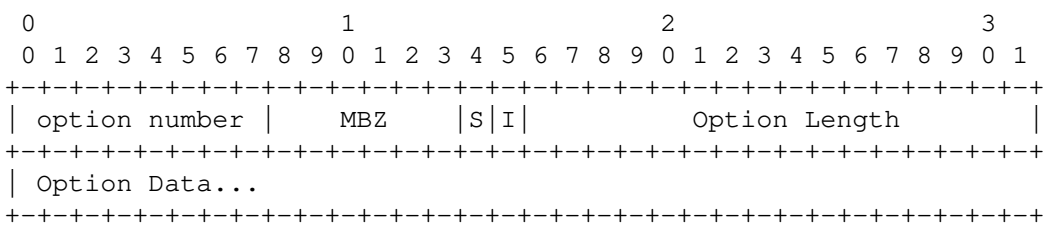
There is no type-specific header for Report and Leave messages; the data field starts immediately after the checksum.

2.5. Data

The remainder of the packet consists of a series of groups and options. Each field in the rest of the packet is either a group address:



or an option header:



The two forms may be told apart because option numbers are assigned in the range [0,223], the first byte of an IPv4 group address is in the range [224,239], and the first byte of an IPv6 group address is 255.

2.5.1. Data Description

2.5.1.1. Group Address

The group address being reported or queried.

2.5.1.2. Option Number

The number of the option.

2.5.1.3. MBZ

Must be zeroed on transmission and ignored on reception.

2.5.1.4. I

Ignore this packet or group for group membership purposes if the option is not recognized.

2.5.1.5. S

Ignore this packet or group for report suppression purposes (on Reports or Leaves) or for reply purposes (on Queries) if the option is not recognized.

2.5.1.6. Option Length

The number of words, excluding the initial word, of option data following the option header.

2.5.1.7. Option Data

Option Length words of data.

2.5.2. Option Processing

Options which precede any group addresses are called Global Options. Options which follow a group address are associated with that group address and are called Group Options. There are two bits describing how to handle unsupported options.

2.5.2.1. The S bit

The S bit is used when processing Queries, Reports and Leaves by interior routers. Groups with options attached should be ignored as if they were not present if there are unrecognized Group Options with the S bit set. Packets with Global Options should be ignored as if they were not received if there are unrecognized Global Options with the S bit set.

2.5.2.2. The I bit

The I bit is used when processing Reports and Leaves by border routers. Groups with options attached should be ignored as if they were not present if there are unrecognized Group Options with the I bit set. Packets with Global Options should be ignored as if they were not received if there are unrecognized Global Options with the I bit set.

2.5.3. Defined Options

2.5.3.1. Padding (option #0)

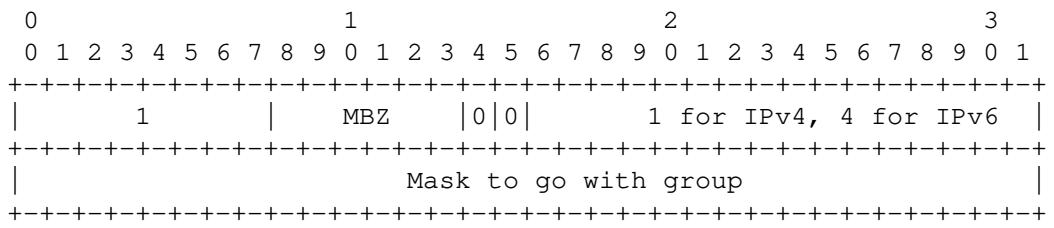
This option need not be handled specially by option parsers; it may be left as an unrecognized option. The S and I bits are both 0, so failing to recognize this option does not affect the processing of the packet. The length field may be 0, meaning there are 0 additional words of data associated with the option. A non-zero length field may be used with the padding option if additional padding is required.

Routers MUST interpret the S and I bits of Padding options as though the option is not supported.

2.5.3.2. Group masks accepted/present (option #1)

This option may be used as a global option on a Query, to indicate that all border routers understand the group mask option in Report and Leave messages. This option MUST only be sent when the Querier knows that all border routers support it; in general this can only be by manual configuration. In this use, the I and S bits are off.

When the most recent Query message contained the Group masks accepted global option, a router may attach a group masks present option to any group in its Report or Leave messages. This option contains the following data:



The data portion is a bitwise mask, to be applied to the group to create a group range. (XXX should it be a mask length? Are there still proposed address allocation schemes which use noncontiguous masks?)

This usage also has the S and I bits turned off.

2.5.3.3. Unicast-reply (option #2)

This option has the S and I bits turned off. If a query is received with this option, the reply should be unicasted to the source of the query. If the option carries a unicast address, it is the address to be unicasted to.

If a unicast address is specified in the option, the option MUST be ignored if the packet is not fully authenticated using IPSEC[Atki95].

3. Message Descriptions

3.1. Domain-Wide Query

A Domain-Wide Query is sent periodically by one of the domain's border routers. The default period is 5 minutes, and MUST be configurable. Domain-Wide Query messages are sent to the well-known Domain-Wide Query multicast group (224.0.255.254). This group is in the range of addresses that are scoped to a single domain, 224.0.255.0 through 224.0.255.255.

A Domain-Wide Query with no additional data is a request for knowledge of all multicast group memberships in the domain. The Domain-Wide Query may be restricted by including groups or options in the data portion of the packet. If group addresses or DWR options are specified in the packet, the Query is restricted to those groups or other options as specified in the packet.

3.2. Domain-Wide Report

A Domain-Wide Report is sent by a router in response to a Domain-Wide Query message, or in response to the appearance of a new group member in the domain. The latter messages are called "Unsolicited" Domain-Wide Reports.

A Domain-Wide Report message includes a list of group memberships and other options in the additional data portion of the packet.

3.3. Domain-Wide Leave

A Domain-Wide Leave is sent by a router when it knows that there are no more members in the domain of a group or groups. The group or groups listed in the additional data portion of the packet are considered by the border routers to have no more members.

3.4. Non-Authoritative Domain-Wide Leave

A Non-Authoritative Domain-Wide Leave is sent by a router when it knows of no more members of the group but cannot be sure there are no more members in the domain. Reception of a Domain-Wide Leave causes the elected border router to send a Domain-Wide Query message for the group(s) mentioned in the Non-Authoritative Domain-Wide Leave message.

4. Interior Router Behavior

4.1. General Behavior

This section describes the general behavior of interior routers, or of proxies running inside domains. Some of these behaviors may be optimized when running multicast routing protocols with more centralized knowledge of group memberships inside the domain; these optimizations will be described later.

If a router has not yet been upgraded to perform domain-wide reports, a proxy may be placed on each of its connected networks. This proxy must participate in the network's group membership protocol (e.g. IGMPv2[Fenn97]). For example, it may perform only the duties of a Non-Querier router in IGMPv2, which allow it to passively learn all of the group members on a network. The proxy can then respond to Domain-Wide Query messages just as the interior router would.

4.1.1. Reception of Queries

All routers in the domain join the Domain-Wide Query well-known multicast group. Upon reception of a Domain-Wide Query message, a router sets a timer to a value randomly chosen from the range (0, Response time] as specified in the packet. The Data section of the Query should be saved to be used when the timer expires.

4.1.2. Transmission of Reports

Upon the expiration of a Domain-Wide Query timer, a router assembles a packet containing the list of group memberships known to this router via IGMP or other mechanism, excluding those that were suppressed by previous reports, and sends this message to the Domain-Wide Report well-known multicast group (224.0.255.253). If the Domain-Wide Query contained a list of groups or options, the Report should be restricted to those groups in the list in the Query message.

4.1.3. Reception of Reports

All routers in the domain join the Domain-Wide Report well-known multicast group in order to perform suppression, as follows. Upon reception of a Domain-Wide Report message, a router processes the list of groups in the message. If the packet contains unrecognized global options, the packet should be dropped and not processed if any of the unrecognized options have their S bit set. For each group, if the group has unrecognized options, the group should be skipped if any of the unrecognized options have their S bit set. Otherwise, if the router's Domain-Wide Query timer is running, it SHOULD mark the group as having been suppressed and SHOULD NOT report it when the Domain-Wide Query timer expires.

Routers MAY record the existence of this group membership in the domain to be used for future suppression. This record MUST time out after [Query Interval] * [Robustness Variable], and MUST be canceled by reception of a Domain-Wide Leave or Non-Authoritative Domain-Wide Leave message mentioning this group.

4.1.4. Reception of Leaves

Upon the reception of a Domain-Wide Leave, a router should process the list of groups in the message. For each group, if the group has unrecognized options, it should be skipped if any of the unrecognized options have their S bit set. Otherwise, the router should remove its record of the existence of another group membership in the domain.

4.1.5. Transmission of Leaves

A router sends a Non-Authoritative Leave when it no longer knows of any members of the group. This message MAY be suppressed if some other router was the last to report group membership with a DWR.

4.2. Optimizations

In explicit group membership protocols like PIM, CBT and MOSPF, there is a set of routers smaller than "all routers in the domain" which knows of group memberships in the domain. This section describes the optimizations possible when running a protocol like this.

In PIM and CBT, only RP's and Cores must participate. MOSPF is a special case, in that all routers in the MOSPF domain know of all group memberships in the domain. In this case, the DWR protocol may degenerate to a virtual protocol run entirely inside the elected border router.

4.2.1. Reception of a Query Message

Only participating interior routers must join the Domain-Wide Query well-known multicast group.

4.2.2. Transmission of a Report Message

Report messages contain all memberships that this router knows about (e.g. in MOSPF, it's all memberships in the domain; in PIM, it's all groups that for which this router is the RP).

4.2.3. Reception of a Report Message

If there is no overlap of the groups being reported by each participant, the interior routers need not join the Domain-Wide Report well-known multicast group so will not receive Report messages. E.g. if R1 and R2 each handle one half of the multicast group address space, there is no need for either of them to join the Domain-Wide Report group.

4.2.4. Reception of a Leave Message

As with Reports, if there is no overlap, the interior routers need not join the DWR group so will not receive these messages.

4.2.5. Transmission of a Leave Message

If it is possible to know when the last member in the domain goes away, routers SHOULD send authoritative Domain-Wide Leave messages, instead of Non-Authoritative Domain-Wide Leave messages.

5. Unsolicited messages

When a new group member appears in the domain, a Report message SHOULD be transmitted (called an Unsolicited Report). Interior routers MAY track the presence of group members inside the domain; a router which is doing this SHOULD suppress its unsolicited Report if it knows of another group member inside the domain.

6. Border Router Behavior

6.1. Querier Election

All border routers should join the Domain-Wide Queries well-known multicast group, in order to perform Querier Election. All routers start up thinking they are the elected Querier. If a router hears a DWQ from a router with a lower IP address, it elects that router as Querier. If a router has not heard a DWQ from the elected Querier in [Querier's Query Interval] * [Querier's Robustness Variable] + [Querier's Response Interval], it assumes the role of Querier. It is recommended to have an IPSEC[Atki95] relationship among the domain border routers so that Querier Election can not be fouled by a forged packet.

6.2. Send Periodic Queries

The elected border router sends periodic Queries once every [Query Interval]. These Queries include the router's Query Interval and Robustness Variable. The Response Interval should be set to [Normal Response Interval].

6.3. Reception of Non-Authoritative Leave

Upon reception of a Non-Authoritative Leave, the elected Querier sets a group membership timeout timer to [Robustness Variable] * [Response Interval] + [Round Trip Delay], and sends a group-specific Query, listing all groups in the Non-Authoritative Leave message. The Response Interval should be set to [Fast Response Interval]. Until a response is heard for each listed group, the Query should be retransmitted once every [Fast Response Interval] for a total of [Robustness Variable] transmissions. The Querier MUST wait an additional

[Round Trip Delay] after the final [Fast Response Interval] for reports before assuming that there are no members present in the domain.

6.4. Reception of Group-Specific Queries

Upon reception of a Group-Specific Query, non-Querier routers MUST set a group membership timeout timer to [Robustness Variable] * [Response Interval] + [Round Trip Delay]. If this timeout occurs without receiving a Report for the listed groups, the group membership record is removed.

6.5. Reception of Reports

Upon reception of a Domain-Wide Report message, all border routers set a group membership timer for each group mentioned in the Report. This timer's value is set to [Querier's Query Interval] * [Querier's Robustness Variable] + [Querier's Response Interval] * 2. The Querier's Query Interval, Querier's Response Interval and Querier's Robustness Variable are remembered from the last General Query received from the Querier. This timer is refreshed by reception of further messages.

6.6. Request Unicast Replies

If a Border Router wishes to reduce the amount of DWR multicast traffic in a domain, it may add the "Reply via Unicast" option to its DWQ's. This has the advantage of reducing the amount of state kept inside the domain for forwarding packets destined to the DWR-reply multicast group, at the cost of eliminating suppression. The border router must multicast DWR's summarizing the replies it got via unicast to the DWR-reply multicast group at the end of the response interval, in order to share membership information with all routers. This summary MUST contain a global padding option with its S bit set to 1, to prevent suppression of real reports.

7. List of timers and tunable values

7.1. Robustness Variable

The Robustness Variable allows tuning for the expected packet loss in a domain. If transmission inside a domain is expected to be lossy, the Robustness Variable may be increased, at the cost of increased latency in determining failures. The DWR protocol is robust to ([Robustness Variable] - 1) packet losses. The Robustness Variable MUST NOT be zero, and SHOULD NOT be one. Default value: 2

7.2. Query Interval

The Query Interval is the interval between General Queries sent by the domain-wide Querier. Default value: 5 minutes

7.3. Normal Response Interval

The Normal Response Interval is the Response Time inserted into the periodic General Queries. Default: 60 seconds

By varying the [Normal Response Interval], an administrator may tune the burstiness of DWR messages in the domain; larger values make the traffic less bursty, as host responses are spread out over a larger interval. The number of seconds represented by the [Normal Response Interval] must be less than the [Query Interval].

7.4. Fast Response Interval

The Fast Response Interval is the Response Time inserted into Group-Specific Queries in response to Non-Authoritative Leave messages, and is also the time between the Group-Specific Query messages. Default: 1 second

This value may be tuned to modify the "leave latency" of the domain. A reduced value results in reduced time to detect the loss of the last member of a group.

7.5. Round Trip Delay

The Round Trip Delay is the worst-case round trip time through the domain. This is used to ensure that group membership is not lost due to a small Fast Response Interval and a large round trip delay through the domain. This value must be manually configured. Default: 100ms.

IGMPv2 ignores end-to-end message delay, assuming that this delay is negligible. Although the DWR protocol is very similar to IGMPv2, the reality is that end-to-end round trip delays can be very different on LANs vs. in a routing domain. On a LAN, the round trip delay is generally dwarfed by the IGMPv2 response interval. Within a domain, the opposite may be true, so it's important for the protocol to acknowledge that.

8. Message destinations

This information is provided elsewhere in the document, but is summarized here for convenience.

Message Type	Destination Group
General Query	Domain-wide Query group (224.0.255.254)
Group-specific Query	Domain-wide Query group (224.0.255.254)
Report	Domain-wide Report group (224.0.255.253)
Leave	Domain-wide Report group (224.0.255.253)
Non-Authoritative Leave	Domain-wide Report group (224.0.255.253)

9. Miscellaneous issues

9.1. Elect Reporters?

In order to help with suppression in a domain, a Querier might choose to elect a "Designated Reporter" for a group for a certain duration. The Designated Reporter is the only router which should send Reports for the designated group(s) for the designated time. All others should act as though they have been suppressed for the designated time. The Querier should cancel a router's Designated Reporter status when that router sends a Leave message, or when it hasn't heard a reply from that router for <N> Query intervals. When canceling a router's Designated Reporter status, a Querier should send a Group-Specific Query for the group(s) in question and can optionally elect one of the responders as the new Designated Reporter.

Designated Reporters are only required when using unicast replies; when using multicast replies, routers may keep track of the presence of other reporters in the natural course of handling suppression.

10. Acknowledgments

The ideas of unicasting DWR replies and of electing a "designated reporter" came from a discussion on the IDMR mailing list with Jeffrey Zhang and Yunzhou Li of Bay Networks.

11. Security Considerations

Many same as IGMPv2[Fenn97]

11.1. Unicast Responses

Sending a DWQ requesting a unicast response can cause many DWR's to be unicasted to the sender. Requiring IPSEC authentication on the DWQ only if it requests unicast to a different address may not be strong enough - for example, someone at the other end of a slow link may swamp the link by sending a DWQ. (At the same time, someone at the other end of a slow link may swamp the link by sending a multicast ping...)

12. References

Atki95 Atkinson, R., ''IP Authentication Header'', RFC 1826, August 1995.

Brad88 Braden, B., D. Borman, C. Partridge, "Computing the Internet Checksum", RFC 1071, ISI, September 1988.

Bradner97 Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119/BCP 14, Harvard University, March 1997.

Fenn97 Fenner, W. ‘‘Internet Group Management Protocol, Version 2’’, RFC2236, Xerox PARC, November 1997.

Estr98 Estrin, D., D. Meyer, D. Thaler, ‘‘Border Gateway Multicast Protocol (BGMP): Protocol Specification’’, Work In Progress, March 1998.

Thya95 Thyagarajan, A. and S. Deering, ‘‘Hierarchical Distance-Vector Multicast Routing’’, Proceedings of ACM Sigcomm, September 1995.

13. Author’s Address

William C. Fenner
Xerox PARC
3333 Coyote Hill Road
Palo Alto, CA 94304
Phone: +1 650 812 4816
Email: fenner@parc.xerox.com