Internet Engineering Task Force
INTERNET-DRAFT
draft-ietf-idmr-msnip-00.ps

IDMR WG
Bill Fenner/AT&T
Hugh Holbrook/Cisco
Isidor Kouvelas/Cisco

23 February 2001
Expires: August 2001

Multicast Source Notification of Interest Protocol (MSNIP)

Abstract

This document discusses the Multicast Source Interest Notification Protocol (MSNIP). MSNIP is an extension to IGMPv3 [1] that provides membership notification services for sources of multicast traffic.

Table of Contents

1. Introduction

The Multicast Source Notification of Interest Protocol (MSNIP) is an extension to version 3 of the Internet Group Membership Protocol (IGMPv3 [1]) that enables multicast sources to avoid the work of sending packets when there are no receivers. The scenario may be one where there is a server sourcing a large number of multicast flows from a much larger pool of potential multicast flows that map onto separate multicast group addresses. If the source were to continuously transmit data on all the groups that could potentially have receivers, significant resources would be wasted in the system itself as well as the first-hop link. Using a higher level control protocol to determine the existence of receivers for particular flows is not practical as there may be many potential receivers.

Information on which multicast groups have receivers for a particular sender is typically maintained by the multicast routing protocol on the first hop router to the source. MSNIP uses this information to notify the application in the sending system of when it should start or stop transmitting. This is achieved without any group-specific state on the first-hop router for potential sources without receivers.

Many currently deployed multicast routing protocols, require data from an active source to be propagated past the first-hop router before information on the existence of receivers becomes available on the first-hop. All dense-mode protocols fall under this category as well as sparse-mode protocols that use shared trees for source discovery (such as PIM-SM [3]). In order to provide receiver interest notification for such protocols, the default mode of operation would require that the system always transmits on all potential groups and the first-hop routers prune the traffic back.

In order to avoid this complication MSNIP only supports sparse-mode multicast routing protocols that build source-specific trees (such as PIM-SSM [3]). With such protocols, the first-hop router can obtain information on the existence of receivers without forwarding any data from the source. Support for this type of protocols covers the majority of applications that MSNIP is targeted at.

2. API for Requesting Membership Notification

Applications within an IP system that wish to source multicast packets and are interested in being notified on the existence of receivers must register with the IP layer of the system. MSNIP requires that within the IP system, there is (at least conceptually) an Application Programming Interface or API that can be used to register with the IP layer for such notifications. A system's IP API must support the following operation or any logical equivalent:

```
IPMulticastsSourceRegister (socket, interface, multicast-address)
IPMulticastsSourceDeregister (socket, interface, multicast-address)
```

In addition the application must provide the following API for receiving notifications from the IP system:

```
IPMulticastSourceStart (socket, interface, multicast-address)
IPMulticastSourceStop (socket, interface, multicast-address)
```

where:

socket

is an implementation-specific parameter used to distinguish among different requesting entities (e.g., programs or processes) within the system; the socket parameter of BSD Unix system calls is a specific example.

interface

is a local identifier of the network interface on which the application wishes to transmit data to the specified multicast address. An implementation may allow a special "unspecified" value to be passed as the interface parameter, in which case the request would apply to the "primary" or "default" interface of the system (perhaps established by system configuration). If transmission to the same multicast address is desired on more than one interface, IPMulticastSourceRegister is invoked separately for each desired interface.

multicast-address

is the IP multicast address to which the request pertains. If transmission to more than one multicast address on a given interface is desired, IPMulticastSourceRegister is invoked separately for each desired multicast address.

3. Protocol Description

Like IGMP, MSNIP is an asymmetric protocol specifying different behaviors for systems wishing to source traffic and for multicast routers. Note that a system may perform at the same time more than one of the above functions. An example is a router that wishes to source traffic.

3.1. Group Range Negotiation

With current multicast deployment in the Internet, different multicast protocols coexist and operate under different parts of the multicast-address space. Multicast routers are consistently configured with information that maps specific group ranges to multicast protocols. Part of this configuration describes the subset of the multicast address space that is used by source-specific multicast [4].

The default behavior of IP multicast, without MSNIP, is that a multicast application must assume that there are multicast receivers present in the network. In order to allow hosts to avoid transmitting when there are no receivers, MSNIP communicates a range of MSNIP managed groups to source systems.

This task is left up to the IGMP querying router. The querier periodically multicasts a MSNIP Group Map message containing the definition of the group ranges over which MSNIP operates. This destination address of the Group Map message is the ALLSYSTEMS group. The Group Map message is sent every [Group Map Interval] seconds. The message also contains a holdtime which is set to [Group Map Holdtime] and instructs IP systems to maintain the group mapping state for the specified holdtime.

In addition Group Map messages are sent when either:

- the IGMP querier on a link receives an Interest Solicitation message (described in section 3.2) from an IP system that was not previously registered with MSNIP or was registered with a different GenID (see section 6.2.
- the configured ranges over which MSNIP operated change.

When either of the above two events occur the querier initiates a set of [Robustness Variable] Group Map messages.

Upon receipt of a Group Map message, an IP system builds or updates a set of group range records as follows. For each group range present in the message, the system either creates or updates a record of the form:

(interface, group prefix, group mask)

Where interface is the interface the Group Map message was received on and prefix and mask are the definition of the group range.

Each previously existing range record with an interface equal to the interface the message was received on which is not reported in the received message is deleted.

In addition to the group range records, the IP system maintains a holdtime timer associated with the interface which is initialised to the holdtime in the received message. If the timer expires before the receipt of the next Group Map message, all range records related to the interface are deleted.

3.2. Host Interest Solicitation

Hosts that wish to be managed by MSNIP periodically transmit an Interest Solicitation message. This message is multicast with a group destination address of ALL_IGMPv3_ROUTERS (224.0.0.22) and is transmitted every [Interest Solicitation Interval] seconds. The Interest Solicitation message contains a holdtime which is set to [Interest Solicitation Holdtime] and instructs the routers to maintain MSNIP state for this system for the specified period. A generation ID is also included in the Interest Solicitation message to provide support for routers to detect system crashes (see section 6.2).

When an IP system first comes up it transmits [Robustness Variable] Interest Solicitation messages randomly spaced over [Initial Interest Solicitation Interval] seconds.

All MSNIP capable routers on the link that receive an Interest Solicitation message from an IP system, maintain a system interest record of the form:

(IP system address, holdtime timer)

Each time an Interest Solicitation message is received from the IP system, the holdtime timer is reset to the holdtime in the received message. In addition the router responds to the solicitation message with a Receiver Membership Report message described in section 3.3. The message contains a TRANSMIT group record for each of the group addresses within the MSNIP managed range for which the routing protocol indicates there are receivers for this source system.

When the holdtime timer of a specific system interest record expires, the record is deleted.

3.3. Router Receiver Membership Reports

Receiver Membership Report messages are used by routers to communicate the receiver membership status of particular groups to a specific IP system. Each message contains a list of group records of either TRANSMIT or HOLD type that instruct a system to respectively start or stop sending traffic on this link to the specified group address. Receiver Membership Report messages are unicast to the target system.

In addition to the receipt of an Interest Solicitation message, routers send Receiver Membership Reports to IP systems when the receiver membership status reported by the multicast routing protocol changes for a specific source and group. Such reports are only sent if the group is managed by MSNIP and the router has a system interest record created by a previously received Interest Solicitation message with a IP system address equal to the source address. If the source group pair satisfy these conditions then [Robustness Variable] Receiver Membership Reports are sent out within [Unsolicited Membership Report Interval] seconds. If during the [Unsolicited Membership Report Interval] an additional membership change occurs for the same group and source system, transmission of any related pending Receiver Membership Report messages is canceled and a new set of [Robustness Variable] messages is scheduled.

When an IP system receives a Receiver Membership Report message, for each of the TRANSMIT group records listed in the message it creates or updates a group record of the form:

(interface, router address, group address, holdtime timer)

Where the interface is the one the message was received on, the outer address is obtained from the source address on the IP header of the received message and the holdtime timer is set to [Interest Solicitation Holdtime] seconds.

For each group HOLD group record present in the message, the system deletes the relevant group record from its state.

4. Application Notification

This section describes how protocol events trigger notifications to source applications within an IP system.

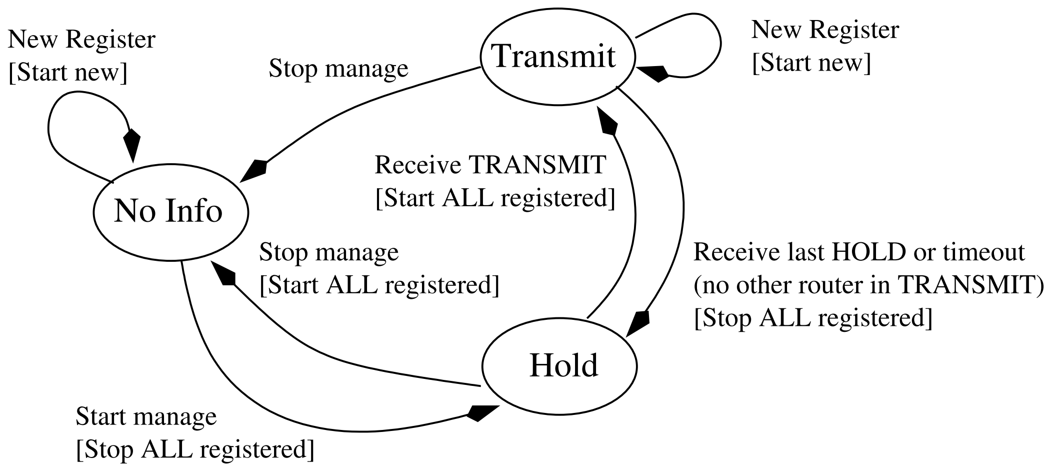

Figure 1: Per Interface (I) and Group (G) state machine at an IP system

In tabular form, the state-machine is:

Prev State	Event				
	New Register	**Start Manage**	**Stop Manage**	**Recv TRANSMIT**	**Recv last HOLD or timeout**
No Info	Start new	-> Hold Stop ALL registered	-	-	-
Hold	-	-	-> No Info Stop ALL registered	-> Transmit Start ALL registered	-
Transmit	Start new	-	-> No Info	-	-> Hold Stop ALL registered

The events in state machine above have the following meaning:

New register
A new application has registered through a call to IPMulticastsSourceRegister for this G and I.

Start manage
We received a Group Map packet on I that changed the status of G from a non-managed to a MSNIP managed group.

Stop manage
We received a Group Map packet on I that changed the status of G from a MSNIP managed to a non-managed group or the mapping that caused this group to be managed expired.

Receive TRANSMIT
We received a Receiver Membership Report on I that contains a TRANSMIT group record for G.

Receive last HOLD or timeout
We either received a Receiver Membership Report on I that contains a HOLD group record for G or a TRANSMIT group record gor G on I expired and there are no other transmit records for G on I. This means that there are no routers left wishing to receive traffic from this system to group G on I.

The state machine actions have the following meaning:

Start new
Send an IPMulticastSourceStart notification to the application that just registered for this G and I.

Start ALL registered
Send an IPMulticastSourceStart notification to all applications that are registered for this G and I.

Stop ALL registered
Send an IPMulticastSourceStop notification to all applications that are registered for this G and I.

5. Router Processing

This section describes the per-source system tracking that takes place within a router.

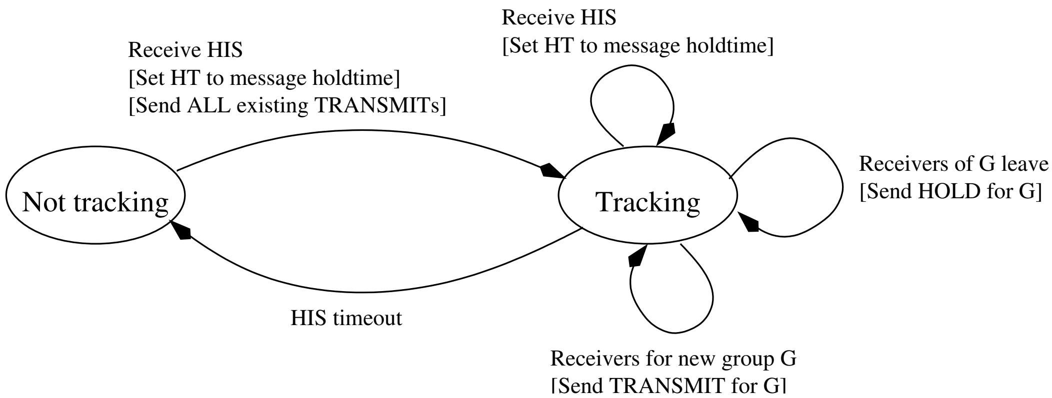

Figure 2: Per IP source system (S) state machine at a router

In tabular form, the state-machine is:

Prev State	Event			
	Receive HIS	**HIS timeout**	**Receivers for new group G**	**Receivers of G leave**
Not tracking	-> Tracking Set HT to message holdtime Send ALL existing TRANSMITs	-	-	-
Tracking	Set HT to message holdtime	-> Not tracking	Send TRANSMIT for G	Send HOLD for G

The events in state machine above have the following meaning:

Receive HIS
The router has received a Host Interest Solicitation from S.

HIS timeout
The holdtime timer (HT) associated with S has expired.

Receivers for new group G
The routing protocol has informed MSNIP that it now has receivers for the MSNIP managed group G and system S.

Receivers of G leave
The routing protocol has informed MSNIP that all receivers for the MSNIP managed group G and system S have left.

The state machine actions have the following meaning:

Set HT to message holdtime
The holdtime timer associated with S is restarted to the value of the holdtime in the received Host Interest Solicitation message.

Send ALL existing TRANSMITs
The router builds and transmits Receiver Membership Reports to S that contain a TRANSMIT group block for each of the MSNIP managed groups that have receivers for S.

Send TRANSMIT for G
The router builds and transmits a Receiver Membership Report to S that contains a TRANSMIT group block for the group G.

Send HOLD for G
The router builds and transmits a Receiver Membership Report to S that contains a HOLD group block for the group G.

6. Message Formats

Like all other IGMP messages, MSNIP messages are encapsulated in IPv4 datagrams, with an IP protocol number of 2. Every MSNIP message described in this document is sent with an IP Time-to-Live of 1, and carries an IP Router Alert option [RFC-2113] in its IP header.

There are three IGMP message types of concern to the MSNIP protocol described in this document:

Type Number (hex)	Message Name
0x23	Group Map
0x24	Interest Solicitation
0x25	Receiver Membership Report

6.1. Group Map Packet

A Group Map packet is periodically multicast by the IGMP querier to declare the multicast group ranges manages by MSNIP. The Group Map message is multicast with a group destination address of ALL_IGMPv3_ROUTERS (224.0.0.22).

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Type = 0x23  |  Group Count  |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Holdtime                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Group-Prefix-1                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Mask-Len-1   |                   Reserved                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               .                               |
|                               .                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Group Count

The number of group range records present in this message. Note that the actual maximum number of group ranges that can be reported is limited by the maximum size of an IP packet. As each Group Map packet replaces the mapping at a receiving system, a router cannot split a group mapping into more than one Group Map packets.

Checksum

The Checksum is the 16-bit one's complement of the one's complement sum of the whole MSNIP message (the entire IP payload). For computing the checksum, the Checksum field is set to zero. When receiving packets, the checksum MUST be verified before processing a packet.

Holdtime

The amount of time a receiving system must keep the group map state alive, in seconds.

Group-Prefix-1

The group prefix of the first group record present in this message.

Mask-Len-1

The mask length for the group range in the first group record present in this message.

Reserved

Transmitted as zero. Ignored upon receipt.

6.2. Interest Solicitation Packet

A Interest Solicitation packet is periodically multicast by MSNIP capable systems to declare interest in Receiver Membership Reports from multicast routers on the link. The Interest Solicitation message is multicast with a group destination address of ALL_IGMPv3_ROUTERS (224.0.0.22).

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Type = 0x24  |    Reserved   |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Holdtime            |             GenID             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Reserved
Transmitted as zero. Ignored upon receipt.

Checksum
Calculated as for Group Map packet.

Holdtime
The amount of time a receiving router must keep the system interest state alive, in seconds.

GenID
Generation ID of the IP system. A number that is selected randomly for each of the [Robustness Variable] initial Interest Solicitation messages when the system comes up and afterwards remains fixed to the value used in the last of the initial messages throughout the system lifetime. The GenID is used by routers to detect system crashes.

6.3. Group Membership Report Packet

A Group Membership Report packet is unicast by first-hop multicast routers and targeted at potential sources to inform them of the existence or not of receivers for the listed multicast groups.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Type = 0x25  |  Group Count  |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Record-Type-1 |                   Reserved                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Group-Address-1                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Group Count
The number of group records present in this message.

Checksum
Calculated as for Group Map packet.

Record-Type-1
The type of the first group record in this message. Valid values are:

Record Type	Description	Value
TRANSMIT	Request to start transmitting group	1
HOLD	Request to stop transmitting group	2

Reserved
Transmitted as zero. Ignored upon receipt.

Group-Address-1
The multicast address of the first group record in the message.

7. Constants Timers and Default Values

Robustness Variable
The Robustness Variable allows tuning for the expected packet loss on a network. If a network is expected to be lossy, the Robustness Variable may be increased. MSNIP is robust to (Robustness Variable - 1) packet losses. The Robustness Variable MUST NOT be zero, and SHOULD NOT be one. Default: 2

Group Map Interval
The interval used by the IGMP querier between transmissions of Group Map messages. Default: 60 secs

Group Map Holdtime
The interval inserted in Group Map messages that indicates to systems how long they should use the included mapping information before they time it out. This MUST be ((the Robustness Variable) times (the Group Map Interval) plus (one second)).

Interest Solicitation Interval
The interval used by MSNIP capable systems between transmissions of Interest Solicitation messages. Default: 60 secs

Interest Solicitation Holdtime
The interval inserted in Interest Solicitation messages by systems to instruct routers how long they should maintain system interest state for. This MUST be ((the Robustness Variable) times (the Interest Solicitation Interval) plus (one second)).

Initial Interest Solicitation Interval
The interval used by systems to send out the initial Interest Solicitation messages when they first come up. Default: 1 second.

Unsolicited Membership Report Interval
The interval used by routers to send out a set of Membership Report messages when the group membership changes for a specific system. Default: 1 second.

8. Todo list...

- Add security considerations section.
- If application ignores or does not ask for notification in managed range host OS should filter packets.
- Maybe provide masks for registration / notification API.
- Case where host and app starts before MSNIP range is available.
- When we hear out-of-order IGMP query resend interest registration.
- Only send interest registration when application is interested. This is not possible if we do host filtering...
- Maybe add API to ask the kernel for the state of a particular group. bool IpMulticastSourceHasInterest (socket, interface, multicast-address).
- Add GenID changes to router FSM.

9. Authors' Addresses

Bill Fenner
AT&T Labs - Research
75 Willow Road
Menlo Park, CA 94025
fenner@research.att.com

Hugh Holbrook
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
holbrook@cisco.com

Isidor Kouvelas
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
kouvelas@cisco.com

10. Acknowledgments

The authors would like to thank Jon Crowcroft for his contribution to this specification.

11. References

[1] B. Cain, S Deering, W. Fenner, I Kouvelas, A. Thyagarajan, "Internet Group Management Protocol, Version 3", Work In Progress, <draft-ietf-idmr-igmp-v3-05.txt>, 2000.

[2] S. Kent, R. Atkinson, "Security Architecture for the Internet Protocol.", RFC 2401.

[3] B. Fenner, M. Handley, H. Holbrook, I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", Work In Progress, <draft-ietf-pim-sm-v2-new-01.txt>, 2000.

[4] Z. Albanna, K. Almeroth, D. Meyer, "IANA Guidelines for IPv4 Multicast Address Allocation", Best Current Practices, <draft-ietf-iana-IPv4-mcast-guidelines-00.txt>, 2001.