Ethernet Group Membership Protocol (EGMP)
Draft RFC

David R. Cheriton
Stanford University and Granite Systems
cheriton@cs.stanford.edu

Stephen E. Deering
Xerox PARC
deering@xerox.com

Kenneth J. Duda
Granite Systems
kjd@arp.com

October 22, 1995

1 Status of This Memo

This memo is a draft specification of EGMP, a MAC- or datalink-level protocol for explicitly joining and leaving groups corresponding to multicast and unicast addresses on an extended LAN such as switched/bridged Ethernet and other network technologies using Ethernet addresses, such as FDDI. It provides a MAC-level analog to IGMP [6]. That is, EGMP allows the extended LAN to deliver packets addressed to a multicast address to only those LAN segments with endstations that have explicitly joined the group corresponding to that multicast address. Ancillary to the multicast membership, EGMP supports determining the location of endstations corresponding to specific unicast addresses. Distribution of this memo **will** be unlimited. Currently its distribution is restricted until reviewed and revised further.

2 Introduction

Multicast on broadcast networks such as conventional Ethernets is implemented by delivering the packet to each endstation interface and filtering by address at each interface. That is, the network is expected to broadcast the multicast packets and each endstation interface only allows through the multicast packets whose destination addresses match those of the endstation interface multicast filter. Packets to all other addresses are discarded[1]. (This is the normal mode of operation. An endstation interface can also be set to receive all multicast packets, or even all packets, so-called promiscuous mode.) In this basic model, there is no indication to the network whether an endstation is interested in a given multicast packet or not.

In extended LANs consisting of individual LAN segments interconnected by bridges, each bridge forwards a multicast packet on all ports on the spanning tree other than the incoming port. This flooding behavior is required to provide the underlying broadcast distribution described above for the single-segment LAN. This broadcast distribution does not scale well because it loads the whole extended LAN with the sum of the multicast (and broadcast) traffic of all sources on the extended LAN.

[1] In practice, this filtering is imperfect. The typical interface uses a 64-bucket hash which lets through packets that hash to any enabled bucket. With birthday problem collisions, there can be a significant number of unwanted packets coming through the filter.

The introduction of sophisticated Ethernet switches provides the potential for significant scaling of Ethernets. The standard Ethernet interface becomes purely an access protocol. The conventional broadcast implementation is replaced with a switched fabric with higher aggregate bandwidth than that of individual links. However, a key deficiency with the standard Ethernet access protocol is a lack of an indication of the multicast addresses that an endstation wants to receive. The current move to higher-speed Ethernet and the growing use of multicast for video delivery and other high-bandwidth applications further motivates addressing this problem at the present time.

The Ethernet Group Membership Protocol (EGMP) addresses this deficiency. It is a datalink-layer protocol that allows a switch (or bridge) to determine the endstation interfaces that are interested in receiving a particular Ethernet address, both multicast and unicast. EGMP also supports indicating an interest in receiving unicast addresses to allow learning switches and bridges to locate individual endstations. Using this facility in EGMP, a switch can avoid broadcasting packets addressed to unicast addresses that it has not previously located. It also provides a means for a switch to detect whether an endstation with a particular unicast address supports EGMP.

EGMP is also used as an inter-switch protocol to communicate membership in groups corresponding to multicast and unicast addresses between switches in a multi-switch configuration. In essence, a switch acts as a proxy for its connected endstations, joining the groups corresponding to the multicast addresses that its endstations belong to, but effectively joining these groups in other switches. EGMP relies on the standard spanning tree algorithm to avoid packet loops and duplicates in multi-switch configurations. Use of EGMP with other routing protocols is a subject for future study.

EGMP is based in part on the Host Membership Protocol described by Deering [7] (and standardized as IGMP [6]), but operates at the Ethernet level, not the IP level. EGMP provides LAN segment membership within an extended LAN whereas IGMP effectively creates a single membership for entire extended LANs at the routers for each multicast address with local members. EGMP in conjunction with IGMP provides efficient delivery of multicast to hosts interested in particular IP multicast addresses. EGMP can also be used with other network-layer protocols such as IPX, Appletalk and XNS.

In contrast to IGMP version 1, EGMP defines an explicit leave protocol mechanism to reduce "leave latency", as recently added to IGMP version 2 [5]. *Leave latency* refers to the time between when an endstation decides it is no longer interested in being a member of a group (i.e. receiving packets for that group's address) and the packets are no longer being forwarded to this endstation's LAN or LAN segment, assuming this endstation is the last one that was interested in packets to this address on this LAN segment. Low leave latency is important to keep up with this multicast address switching rate and to avoid having memberships persist to old addresses, thereby wasting excessive amounts of bandwidth. For example, a video application that changes memberships rapidly could overwhelm its LAN segment with reception of packets sent to its old memberships that have not been turned off by the switch because of the long leave latency. With the growing use of multicast for high-bandwidth services such as video and distributed virtual reality, and the splitting of the aggregate application bandwidth over multiple multicast addresses, multicast receivers can be expected to switch memberships across multiple multicast addresses fairly quickly. These uses of multicast are expected to become more prevalent in the future making low leave latency of increasing importance. (Note that low EGMP leave latency is of considerable benefit for individual LAN segments even if IGMP leave latency is higher because terminating an EGMP membership still stops the traffic from entering the LAN segment. The switch, which is still receiving the multicast traffic because of IGMP, has a much higher internal bandwidth and is capable of handling a larger number of multicast groups than individual ports and links.

Monitoring IGMP traffic within a switch, a (rejected) alternative to EGMP, has been implemented by at least one vendor. However, this approach does not work in all topologies and can cause holes in the multicast delivery. It also represents a significant layer violation that seems inappropriate to perpetrate.

As a multicast mechanism, EGMP functionality is not provided by InARP [2] or ATMARP [10] as currently specified. The latter leaves open the handling of multicast, and might benefit from the use of EGMP or some extension of EGMP. ATMARP does incorporate the notion of an ARP server, similar in some respects to the role of the interrogator in EGMP. In general, there are some parallels in the structure of the protocols but currently no overlap in functionality.

The authors just recently became aware of the IEEE 802.1 effort to address some of these issues [9]. It is hoped that some merger of these efforts is feasible, rather than two protocols for the same service.

2.1 Transition Plan

Fully deploying EGMP means inverting the conventional multicast delivery from "broadcast on all LAN segments" to "send only if requested". To phase in EGMP, each switch or bridge should allow its ports to operate in one of two modes correspondingly to broadcast mode and EGMP mode, as described in Section 7. In broadcast mode, all multicast packets are forwarded to that port/LAN segment, allowing non-EGMP endstations to receive multicast correctly. When only EGMP devices are connected to a port, the port is placed in EGMP mode so that multicast packets addressed to a given address are only forwarded if they have been requested by joining their respective groups. Switch/bridge vendors can by default ship product with all ports in the broadcast mode so they "plug-and-play" with existing equipment. Customers are motivated to upgrade endstations to support EGMP and to change the ports to EGMP mode as the multicast traffic level increases and the conventional broadcast mode overwhelms LAN segments with unwanted multicast traffic. System vendors are motivated to incorporate EGMP if it is a standard to allow their systems to work well with high-performance switches in demanding applications, such as video conferencing.

2.2 Why Use ONC RPC and GMP?

EGMP is defined in terms of a general-purpose Group Membership Protocol (GMP) defined as a remote procedure call interface, in contrast to conventional MAC and network layer protocols in use in the Internet. The packet formats and basic handling are defined by ONC RPC [17] and XDR [16], treating the GMP procedures as remote procedure calls.

EGMP uses GMP and ONC RPC so as to build on existing technology, to avoid contributing to the growing protocol chaos in the Internet, and to provide the generality that is needed for the future extensions. The basic ONC RPC procedures are described in the appendix.

Considering existing RPC technology, EGMP could be generated using RPC stub generators in the future (although it is not required for an implementation). Also, the RPC transport and authentication mechanisms could be used without changing the basic procedure-specified protocol. For example, EGMP could be extended with backwards compatibility to provide authenticated memberships using the standard ONC RPC security mechanisms. In this case, a client implementation not supporting authentication would simply report an authentication error to the higher-level software when it attempted to use a membership service requiring authentication.

Considering the protocol chaos issue, basing EGMP on GMP provides the potential of using the same basic RPC interface at a number of levels, from datalink level to application-level services. This unity is attractive because a person developing or maintaining a multicast-based service may be forced to understand and perhaps debug multicast protocols and management mechanisms at all the levels. After all, it does have to work at all levels for the application to work.

ONC RPC was chosen because it is the most widely used RPC system and it is relatively simple to describe in packet formats. Although ONC RPC is normally transported over UDP, it can also function using Ethernet packets directly, at least for low-level services such as EGMP that exist at the datalink level. Using RPC at the datalink level is regarded as a "recursive" design [4] in which higher-level protocols are implemented in terms of restricted versions of themselves, rather than as separate ad hoc protocols, as has been the practice to date. As with recursion in general, this recursive structure to the protocols leads to a simpler, more regular design.

As a concession to current Internet practice, EGMP is described at the packet level so it can be implemented directly without knowledge of ONC RPC or XDR and without the use of RPC stub generator tools.

Finally, the cost of using an RPC framework is relatively low. EGMP requires 2 extra packet formats over the single packet format that would be feasible if the design followed the conventional approach as taken by the IGMP Version 1. The use of ONC RPC also adds about 28 bytes to the packet size over an IGMP-like packet design. However, we conjecture that the ability to provide a list of addresses in a single membership packet reduces the number of EGMP packets and the total amount of data sent compared to IGMP version 1 in most network configurations. Moreover, the packet and bandwidth requirements for EGMP is expected to be very low. Finally, the actual number of packets sent is basically the same as a message-based IGMP-like protocol.

The rest of this document describes the RPC basis for EGMP using a generic group membership protocol, the EGMP protocol itself, how it is to be used with IP routers and similar devices, inter-switch operation with EGMP and some suggestions for implementations.

3 GMP: RPC Group Membership Protocol

Group Membership Protocol (GMP) is a general-purpose membership protocol defined as an RPC-implemented interface. We first describe the protocol model and then the specific procedures.

3.1 GMP Model

The basic GMP model is that of a *membership service* provided by one or more servers, allowing clients to join and leave groups. A client requests one or more memberships from the membership server, specifying the groups that it wishes to join. The server can accept or reject the membership request. A client can also request the termination of one or more memberships.

Periodically, the server can propose to terminate one or more or all memberships held by a client or a subset of clients (multicasting the call to that subset in the latter case). The *subset* of member clients typically corresponds to those connected by some common communication mechanism, such as those connected to one port of the switch in the case of EGMP. In response to the server, the clients re-request the memberships to get the memberships extended, if the server is willing and able to do so. By periodically forcing the clients to rejoin, the server gets the clients to reaffirm their interest in the memberships, effectively garbage-collecting memberships that are no longer of interest.

3.2 GMP Procedures

GMP contains the following procedures in its interface.

```
void join( MembershipDesc desc )

void leave( MembershipDesc desc )
```

Both procedures are implemented and exported by a membership server and called by the client. Each EGMP server also implements client EGMP to resend client calls and for inter-switch operation, as described in Section 6.

The semantics of these procedures are as follows:

`void join( MembershipDesc desc )` - Ensure there are memberships for the address(es) described by `desc`. That is, create the memberships if they do not exist or extend existing memberships.

`void leave( MembershipDesc desc )` - Remove the memberships described by `desc` unless another join call is received that requests these memberships for the same client or subset of clients. In the case of EGMP, the subset of clients normally corresponds to those attached to one or more ports of the switch. This is a datagram call.

The format and semantics of the membership descriptions is specific to particular protocols. The format for EGMP descriptions is covered in the next section. The EGMP membership server is normally implemented as part of the Ethernet switch. Other protocols based on GMP can use their own membership servers, with their own membership description formats and separate multicast addresses.

3.3 Role of the RPC System

These procedures are mapped to packets by a remote procedure call system. EGMP uses the ONC RPC and XDR standards to map to packets, except the packet is an Ethernet packet rather than a UDP or TCP packet. Other RPC systems can use the same GMP specification for other membership uses, such as for application server memberships services. The RPC system is also expected to provide authentication and other security services if that is required in the application domain.

4 EGMP Protocol Description

The basic service of EGMP is the protocol to join a group associated with a multicast or unicast address in order to receive packets sent to this address. EGMP also supports *source filtering*, allowing an endstation to ask a switch or bridge to restrict the packets the endstation receives addressed to a specific multicast address to those sent by a specified set of the sources. The joining station can specify the set either by specifying the included sources or the excluded sources. (A similar facility has been proposed for IGMP version 3 [3].)

EGMP uses the Ether type field value allocated for Ethernet-level ONC RPC (yet to be allocated).

In its use as an endstation-to-switch protocol, all EGMP call packets are transmitted to a single well-known (yet to be allocated) Ethernet multicast address referred to as the EGMP address. Call responses (when used) are unicast to the caller. In its use as a switch-to-switch protocol, EGMP call packets are multicast using a separate EGMP inter-switch multicast address (yet to be allocated).

Normally, an Ethernet switch acts as the membership server for EGMP. The switch does not forward (EGMP) packets addressed to this one address to other LAN segments, so endstations on other LAN segments do not receive EGMP traffic not local to their segment. (This differs from the IGMP approach of sending each Report to the multicast address of the group for which it is reporting. Although the EGMP approach means that all endstations on the LAN segment receive every EGMP packet, the level of traffic is not expected to be significant.)

4.1 Membership Descriptor

The general format for an EGMP membership descriptor is

```
+-----+-------+-------+-----------+------+-----------+-----+-----+
| tag | delay | count | EtherAddr0 | prio | EtherAddr1 | prio | ... |
+-----+-------+-------+-----------+------+-----------+-----+-----+
   4      4       4          6         2         6         2
```

where the `tag` field is a 4-byte unsigned value indicating the specific form of leave or join, the `delay` field is a 4-byte unsigned value indicating the time in microseconds for acting on the operation, the `count` is a 4-byte unsigned count of the number of bytes in the description, and each `EtherAddr` field is a 6-byte Ethernet address. The addresses are laid out in big-endian order with the bits in "canonical" (little-endian) order within each byte, i.e. the same order as used in IEEE 802.3 frames. In each call, the addresses specify the memberships of interest.

The 2-byte `prio` field associated with each Ethernet address specifies a priority associated with this membership. A value of 0 means normal delivery. The semantics for other values are yet to be assigned.

The descriptor specifies the packets to deliver based on destination and source address, in different forms depending on the tag value. Each tag name, value and corresponding interpretation is described below.

Null (0) : Terminates a list of descriptors.

Unfiltered(1) : Each Ethernet address specifies a group to join or leave, depending on the call. The `prio` field associated with each address specifies the delivery priority for this address. A delay value of 0 indicates the call is from an endstation whereas a non-zero value indicates a call from a switch with the value indicating the time in microseconds before the operation takes effect, absent any objections. A join call specifying this tag is called a `join-unfiltered`. A leave call specifying this tag is called a `leave-unfiltered`.

IncludedSources(2) : The same as the unfiltered form, but the second and subsequent Ethernet addresses specify sources whose packets to the first address are to be delivered to the group member. The `prio` field for the second and subsequent fields should be zero and is otherwise ignored. A join call specifying this tag is called a `join-including` call. A leave call specifying this tag is called a `leave-including` call. a

ExcludedSources(3) : The same as the unfiltered form, but the second and subsequent Ethernet addresses specify sources whose packets to the first address are not to be delivered to the group member. The `prio` fields for the second and subsequent addresses should be zero and are otherwise ignored. A join call specifying this tag is called a `join-excluding` call. A leave call specifying this tag is called a `leave-excluding` call.

AllMulticast(4) : The join or leave operation applies to all multicast addresses except those specified in the list of Ethernet addresses. This list of addresses is referred to as the *exclusion list*. The `prio` field for each address is zero and is otherwise ignored. A join call specifying this tag is called a `join-all` call. A leave call specifying this tag is called a `leave-all` call.

AllUnicast(5) : The join or leave operation applies to all unicast addresses except those specified in the list of Ethernet addresses. This list of addresses is referred to as the *exclusion list*. The `prio` field for each address is zero and is otherwise ignored. A join specifying this tag is referred to as a *join-all-unicast* call. A leave specifying this tag is referred to as a *leave-all-unicast* call.

A join call other than one of the join-all forms is referred to as a join-specific call. A leave call other than one of the leave-all forms is referred to as a leave-specific call.

4.2 ONC RPC Description

The ONC RPC description of EGMP is two programs, one for server and one for client, namely:

```
typedef unsigned int Time;
typedef opaque EtherAddrList<1460>;
typedef struct {
    unsigned int  tag_;
    Time          delay_;
    EtherAddrList addrList_;
  } Description;

const EGMP_SERVER_PROG = 0x13333333;
const EGMP_CLIENT_PROG = 0x13333334;

#ifdef SERVER_PROG
program EGMP_SERVER {
    version EGMP_SERVER_1 {
        void egmpPing(void) = 0;
        void join( Description ) = 1;
        void leave( Description ) = 2;
      } = 1;
  } = EGMP_SERVER_PROG;

#else

program EGMP_CLIENT {
    version EGMP_CLIENT_1 {
        void egmpPing(void) = 0;
        void join( Description ) = 1;
        void leave( Description ) = 2;
      } = 1;
  } = EGMP_CLIENT_PROG;
```

```
#endif
```

Following ONC RPC conventions, procedure 0 in both server and client is a null "ping" procedure.

"Compiling" this description through the standard `rpcgen` program produces RPC stubs that generate and handle the EGMP packet formats, which are described directly below for version 2 ONC RPC.

4.3 EGMP Packet Formats

The EGMP packet format is the ONC RPC request and reply messages that correspond to the GMP procedure declarations. The following is the single ONC request packet used by EGMP. (ONC RPC uses the term "message" rather than packet, but with EGMP, each message is a separate packet and we are describing packet formats, so we use the term "packet" instead.)

```
|<----- 32 bits ----->|
+---------------------+
|  xid                |
+---------------------+
|  msg_type           |
+---------------------+
|  rpcvers            |
+---------------------+
|  prog               |
+---------------------+
|  vers               |
+---------------------+
|  proc               |
+---------------------+
|  auth_flavor        |
+---------------------+
|  auth_length (0)    |
+---------------------+--------------+
|  tag                |              |
+---------------------+              |
|  delay              |              |
+---------------------+              |
|  length             | Description  |
+---------------------+              |
.  AddrList           .              |
.                     .              |
.                     .              |
+---------------------+--------------+
```

Following ONC RPC conventions, the packet is sent in big-endian network order.

The fields of this call packet format are described below, with all but `delay` and `Description` following standard ONC RPC values.

xid - the transaction identifier, incremented on each EGMP call from each source, starting from 1.

msg_type - 0 for call.

rpcvers - 2, current version of ONC RPC.

prog - 0x1f333333 for client, 0x1f333334 for server (to be assigned for GMP)

vers - 1, first version of EGMP.

proc - corresponding to the 3 procedures of (E)GMP.

- ping - 0
- join - 1
- leave - 2

auth_flavor - the value of 0 for standard `AUTH_NULL`.

auth_length - 0, because of the null authentication.

tag - The tag field is a 32-bit unsigned integer specifying the interpretation of the rest of the fields, as described in Section 4.1

delay - The delay field is a 32-bit unsigned integer specifying the maximum delay in microseconds.

The microsecond granularity is chosen (rather than milliseconds) to allow for fast leave from multicast groups. For example, it is feasible to request leaving a group within a 500 microsecond interval over 100Mb Ethernet using fast machines, and it may be feasible to use even tighter bounds on 1 Gigabit (full-duplex) Ethernet in the future.

length - the length field of an XDR variable-length opaque data type, specifying the number of bytes to follow. It is assumed to be a multiple of 8 bytes, ranging from 0 to the number of 8-byte units that fits in a single Ethernet packet, *i.e.,* $(1500 - 40)/8 = 182$ addresses. If the count is not a multiple of 8, the value is interpreted as rounded down to the next lowest multiple of 8.

etherAddrList - Zero or more Ethernet addresses left-aligned on 8-byte boundaries, padded with a 16-bit priority field in the low-order portion, the number being that which will fit into 8-byte units.

A call can carry multiple membership descriptors, with the last one being terminated by a null descriptor. However, the call message must still fit into a single Ethernet packet. (*An extension to the RPC description is required to provide multiple membership descriptors per call.*

The following is the EGMP return packet format for an accepted call.

```
|<----- 32 bits ----->|
+---------------------+
|  xid                |
+---------------------+
|  msg_type           |
+---------------------+
```

```
|  reply_stat         |
+---------------------+
|  auth_flavor        |
+---------------------+
|  auth_length (0)    |
+---------------------+
|  accept_stat        |
+---------------------+
|  low                |
+---------------------+
|  high               |
+---------------------+
```

The fields of this return packet format are described below,

xid - the transaction identifier, matching the call xid to which this is a return message.

msg_type - 1 for REPLY.

reply_stat - 0 for MSG_ACCEPTED (and otherwise it is a rejected message — see below.)

auth_flavor - the value of 0 for standard `AUTH_NULL`.

auth_length - 0, because of the null authentication.

accept_stat - the standard ONC RPC values, namely `SUCCESS (0)`, `PROG_UNAVAIL`(1), `PROG_MISMATCH` (2), `PROC_UNAVAIL` (3) and `GARBAGE_ARGS`(4).

low,high - are only used with `PROG_MISMATCH` (2) to indicate the low and high versions of the program that are supported by the server, as with standard ONC RPC.

The following is the EGMP return packet format for a rejected call.

```
|<----- 32 bits ----->|
+---------------------+
|  xid                |
+---------------------+
|  msg_type           |
+---------------------+
|  reply_stat         |
+---------------------+
|  reject_stat        |
+---------------------+
|  low                |
+---------------------+
|  high               |
+---------------------+
```

The fields of this return packet format for rejected call are described below,

xid - the transaction identifier, matching the call xid to which this is a return message.

msg_type - 1 for REPLY.

reply_stat - 1 for MSG_DENIED. (and 0 for an accepted message, as described above.)

reject_stat - the standard ONC RPC values, namely `RPC_MISMATCH (0)` or `AUTH_ERROR` (1).

low - used with `RPC_MISMATCH` (0) to indicate the lowest supported RPC version number. With a `reject_stat` of `AUTH_ERROR`, this field is interpreted as the reason for authentication failure, using the standard ONC RPC values of `AUTH_BADCRED` (1), `AUTH_REJECTEDCRED` (2), `AUTH_BADVERF` (3), `AUTH_REJECTEDVERF` (4), and `AUTH_TOOWEAK` (5). (The use of `low` is described in this form rather than introducing yet another return packet format, as would be required to be totally consistent with ONC RPC conventions.) EGMP implementations need not initially support authentication so `AUTH_ERROR` should not occur. However, an implementation should recognize when it does arise and return an indication to the higher-level client software. Authenticated memberships may be required on some networks in the future.

high - only used with `RPC_MISMATCH` to indicate the highest supported RPC version number.

4.4 Basic Protocol Procedures

EGMP follows the basic model of GMP described in Section 3. A client invokes a `join` call on the EGMP membership service by sending to the EGMP multicast address[2]. The server responds with a (unicast) return packet indicating success or else indicating a problem with the membership request[3]. Periodically, the designated EGMP server, the *interrogator*[4], sends an `leave` datagram call to the EGMP multicast address. This call can specify multicast and unicast addresses with or without exclusions and source filtering, using one of the 5 types of membership descriptors. The join call is multicast so that other endstations on the same LAN segment see the call. These other endstations can then skip sending their own join call for the specified addresses so only one endstation on each LAN segment actually sends a join call in the expected case. A `leave` call, also multicast to the EGMP multicast address, notifies the server(s) and the other endstations on the LAN segment of the sending endstation's intent to drop one or more memberships or sources. The membership service drops these memberships for this LAN segment unless one or more join calls for this same address are subsequently received specific to this LAN segment.

If the list of addresses for an EGMP operation is longer than the maximum allowed in the protocol (which corresponds to the maximum that fits in a single Ethernet packet), the caller performs multiple calls in sequence, each containing up to the maximum number of addresses.

The following descriptions elaborate on these basic procedures.

[2] Multicasting the call can be viewed as an alternative to using the multicast address to determine the unicast address of the server and then sending to that unicast address. The low frequency of EGMP calls and the benefits of other clients monitoring these calls makes multicasting the call the preferred approach. However, multicasting these requests also allows there to be multiple servers.

[3] The return packet obviates the need for a client to send multiple packets as in IGMP in the case of a single server. In the expected case, the server responds after one call packet. The responding EGMP server can also send out additional membership calls on the LAN segment to ensure other EGMP servers on this segment received the request.

[4] The EGMP interrogator serves the same function as the interrogator in IGMP. We view the use of the `leave` call in EGMP as a form of interrogating the endstations to determine which memberships are still of interest.

4.4.1 The `join` Call

The EGMP join call is transmitted by a node on its LAN segment, addressed to the EGMP Ethernet multicast address.

The delay field shall be set to 0 meaning an indefinite membership period, namely until an explicit (unchallenged) leave by this member or the next leave-all is invoked by the interrogator or one of the other servers.

The EtherAddrList field contains zero or more valid Ethernet addresses.

The interpretation of the call parameters varies depending on the tag field:

Unfiltered(1) : Join each group specified by an address in the Ethernet address list. The `prio` field associated with each address specifies the delivery priority for this address.

IncludedSources(2) : The same as the unfiltered form but request delivery only for packets sent to the first address that are sent from a source specified as one of the second and subsequent Ethernet addresses, if any. The `prio` fields for the second and subsequent addresse should be zero and are otherwise ignored.

ExcludedSources(3) : The same as the unfiltered form but request delivery only for packets sent to the first address that are sent from a source other than those specified by the second and subsequent Ethernet addresses. The `prio` fields for the second and subsequent addresses should be zero and are otherwise ignored.

AllMulticast(4) : Join all multicast addresses except those specified in the list of Ethernet addresses. The `prio` field for each such address is zero and is otherwise ignored.

AllUnicast(5) : Join all unicast addresses except for those specified in the list of Ethernet addresses. The `prio` field for each such address is zero and is otherwise ignored.

Unicast membership as supported in the protocol arises in three situations. A node invokes this join call: (1) in response to a leave call containing its unicast address or in response to a leave-all-unicast call. and (2) when the node boots (at the point its network interface is ready to receive packets), and (3) to sniff traffic to the specified unicast address. Note that a node does not include its own unicast address in join calls generated in response to a leave-all call.

A join-excluding call can specify a unicast address, typically the unicast address of the endstation requesting the exclusion, followed by one or more unicast addresses for sources to exclude. This call causes the switch to filter out unicast traffic coming from the specified excluded sources when it was sent to this specific endstation.

The source filtering aspect of a join-including or join-excluding call is a *hint* to the switch which it can act on if convenient and supported. Otherwise, the packets from the excluded destinations and sources continue to be forwarded. It is also “soft state” in the sense that the switch can drop the information and simply recreate this state when “reminded” by the endstations. In this sense, switches need to deliver a superset of the packet traffic specified in EGMP; it is optional whether a switch implements the strict subset of delivery that is specified by the client.

A switch needs to join groups that it needs to forward. In the normal case, it invokes a join-all call on each LAN segment from which it forwards packets, specifying in the exclusion list of addresses those that other endstations on the LAN segment have joined. (The latter exclusion means that the endstations continue to notify the switch of their interest, rather than having it overridden by an (unrestricted) join-all call. With this switch behavior, there is an explicit (membership) request, possibly as a join-all, for every address thatis of interest to switches or endstations on the LAN segment. Therefore, an endstation can determine by monitoring the EGMP traffic that there is no interest in a particular multicast and prune traffic to that address back to the endstation, so these packets are not even forwarded onto the LAN segment.

4.4.2 Switch `join` Call Handling

On receiving a `join` call, a switch notes the existence of one or more members on the LAN segment from which the `join` call was received, for each group specified in the call. For each such membership, if the call also specifies desired sources, the switch records these desired sources as well, as specified sources to receive from or as all but the set of specified sources. It need not record the identity of the caller of the `join` call.

The switch then sends a unicast response message to the caller.

The packets forwarded onto a given LAN segment are the union of all those requested by join calls on this LAN segment. In particular, the set of allowed sources for a given address is the union of all sources allowed by all the join calls for this address.

The switch may optionally reinvoke this same `join` call on the same LAN segment to ensure that any other switch on this segment is aware of this membership. This retransmission guards against single packet loss causing another switch to not know of this membership. The number of retransmissions is an administrator controlled parameter.

4.4.3 Endstation `join` Call Handling

On receiving a `join` call, an endstation checks whether it has scheduled a join call for any of the group addresses listed in the received join call, typically in response to a leave call. If so and the source filtering in this call also subsumes that of the local membership, it unschedules the call(s) for each such address.

Optionally, the endstation may support pruning back to the endstation, where it does not send packets destined for a particular multicast address because there are no receivers. In this case, the endstation shall check if it has pruned one or more of the addresses mentioned in the join call back to the endstation and, if so, resume forwarding packets to this address onto the LAN segment.

This completes the processing in the endstation. The endstation does not respond to the call.

4.4.4 The `leave` Call

An `leave` call requests canceling memberships to one or more groups on a LAN segment. It is invoked when an endstation stops listening to a multicast group or when an endstation wants to stop receiving from one or more sources of traffic to a specified multicast address or when an endstation voluntarily disconnects from the network. In this latter case it leaves the group(s) corresponding to its unicast address(es), eliminating the address(es) as a valid destination for packets.

The call is viewed as "requesting" cancelation of a membership because other endstations on the same LAN segment may still need to receive these packets and override the cancellation by joining the group in response to the leave call.

The leave-all call is also sent by the switches periodically to force the endstations to rejoin, thereby allowing it to garbage collect any memberships that are no longer of interest. The leave call reduces the latency to terminate a membership compared to waiting this garbage collection mechanism to cancel the membership.

When a switch ceases to need to forward packets to a given multicast address from a LAN segment, it can send a leave-specific call for this address on this LAN segment. It then also adds this address to the exclusion list that it sends with subsequent join-all requests on this LAN segment.

An endstation shall use 0 as the delay value to indicate the call came from an endstation. When the leave call is invoked by a switch, the delay field logically specifies the time in microseconds within which a join call should be received to counter the reduced delivery of packets proposed by the leave call. A switch invokes leave with a delay that is appropriate to allow endstations to respond with appropriate join calls.

The interpretation of the call parameters varies depending on the tag field:

Unfiltered(1) : Leave each group specified by an address in the Ethernet address list. The `prio` field associated with each address should be zero and is otherwise ignored.

IncludedSources(2) : Request stopping the delivery of packets with destination address as the first address and source address one of the second or subsequent addresses in the address list. The `prio` field for each such addresses should be zero and is otherwise ignored.

ExcludedSources(3) : Request stopping the delivery of packets with destination address as the first address and source addresses other than the second or subsequent addresses in the address list. The `prio` field for all addresses should be zero and is otherwise ignored.

AllMulticast(4) : Request stopping reception of packets to all multicast addresses except those specified in the list of Ethernet addresses. The `prio` field for each such address is zero and is otherwise ignored.

AllUnicast(5) : Request stopping reception of packets to all unicast addresses except those specified in the list of Ethernet addresses. The `prio` field for each such address is zero and is otherwise ignored.

As an optimization in the case of stopping reception of all packets for a particular multicast address, only the endstation that last issued a join call for the group on the LAN segment generates a leave call when leaving the group[5]. If the leaving endstation was not the last one to issue a join call for this group, there is at least one other endstation on this LAN segment interested in this group. This optimization assumes it is unlikely that the last member endstation crashed since the join call, and failed to generate the leave call. In the case of the join caller being the only other member and crashing, the multicast packets for this address are forwarded to this LAN segment unnecessarily until the next leave-all period. This situation is considered unlikely to arise and not a significant problem when it does.

4.4.5 Switch `leave` Call Handling

When an interrogator switch receives a leave call, it schedules a subsequent leave call for the addresses specified by this first call, to be sent after leaveDelay. The value of delay specified in this subsequent call is leaveDelay.

[5]This optimization, due to Rosen Sharma, is also used in IGMP Version 2.

If a join call is received between the time of this first call and leaveDelay microseconds later that requests delivery of packets that are not to be delivered according to the scheduled leave call, the scheduled leave call's descriptor is modified so that is not the case. For example, if the leave call specifies ceasing delivery of packets to a destination m from sources $s1$ and $s2$ and the join call is a join-including call specifying destination m and $s1$, then $s1$ is removed from descriptor of the scheduled leave call.

After leaveDelay microseconds, if the descriptor in the scheduled call is null, the switch deletes the scheduled call and terminates the handling of the leave call. Otherwise, the second leave call is sent and the switch delays for another leaveDelay microseconds.

If a join call is received from that LAN segment whose descriptor conflicts with that specified in the leave call, the membership for each such address is retained for that LAN segment. After leaveDelay microseconds plus some time to allow for packet queuing and processing at the interrogator, the uncontested reduction of the packet delivery is imposed, dropping memberships as well as source filtering as specified.

The number of retransmissions of the leave call before the membership is completely deleted should be user-configurable in the switch.

The value of leaveDelay is recommended to be at least 10 times the maximum packet transmission time, *e.g.,* 1.2 milliseconds for 100 Mb Ethernet and 12 milliseconds for 10 Mb Ethernet. The switch should increase this value if the LAN segment is shared, under heavy traffic, or the leave call specifies a large number of addresses.

The leave call by the interrogator ensures that a single packet loss cannot result in the packet flow for the specified address being stopped on this LAN segment. That is, another member endstation on the same LAN segment sees either the client's leave call packet or the interrogator's leave call unless multiple packet loss occurs. EGMP requires there be at least two leaves before shutdown. However, the switch can skip sending the separate leave call altogether and immediately terminate the packet forwarding if it is certain that the LAN segment contains a single endstation such as when it is explicitly configured with a single endstation on that segment.

A leave-all call removes the join-all membership in the switch for this LAN segment but it does not remove memberships in specific groups specified in the address list. This allows a client or switch to change from a join-all with pruning approach to specific memberships without losing packets.

Switches other than those running the interrogator for a LAN segment perform the same actions as the interrogator except they do not send the second leave call (in response to the endstation's leave call). That is, a non-interrogator switch notes the leave call by the endstation and then, if the following leave call by the interrogator is unanswered after leaveDelay microseconds, stops forwarding traffic onto the LAN segment whose delivery is no longer required. If the interrogator fails to send a leave call, the switch can send a leave call itself, effectively taking over as the interrogator (until silenced by another lower-addressed interrogator).

4.4.6 Endstation `leave` Call Handling

When an endstation on the LAN segment receives a leave call whose descriptor specifies packets that the endstation still wants to receive, it schedules one or more join calls to override those aspects of the leave call.

The leave may specify stopping the delivery of packets:

1. to one or more multicast addresses,

2. to a multicast address from one or more sources,

3. to one or more of unicast addresses[6]

that the endstation is still interested in receiving.

Endstations shall record the delay value used by the interrogator for leave-specific calls as leaveDelay and for leave-all calls as leaveAllDelay. A leave call with a non-zero delay value is assumed to come from the interrogator. These values are used for the delay values in subsequent leave operations, thereby tracking the interrogator's estimate of a suitable delay value.

The node (endstation or switch) behavior uses the same techniques as IGMP to avoid join call implosion, but applied at the datalink layer. That is, in more detail, when a node receives a leave-all call or a leave call designating one or more addresses to which it is interested:

1. It starts a join call timer set to a randomly-chosen value between zero and leaveDelay microseconds if a leave-specific call or else leaveAllDelay if a leave-all call. When the timer expires, a join call packet is transmitted containing the list of the addresses to which the endstation joins that were listed in the leave call and have not already been rejoined by some other endstation since the leave call was received. Thus, join calls from different responding endstations are spread out over a leaveDelay or leaveAllDelay microsecond interval instead of all occurring at once.

2. If a node hears a join call for an address to which it belongs on that network, the node marks this address as rejoined.

The switch does not forward packets destined to the EGMP address between LAN segments of the switch, *i.e.* between different ports. Thus, in the normal case, only one join call is generated per leave call per rejoined multicast address on each LAN segment connected to the switch, namely the one generated by the endstation whose delay timer expires first. (A join call can specify multiple addresses, reducing the call count further.)

The client joins the EGMP group the same as other multicast addresses. This approach avoids special cases in the client driver software. It also means that a switch is signaled on the presence of EGMP clients on the LAN segment by the reception of the join call on the EGMP address.

When an endstation receives a leave call specifying a unicast address that it uses, it sends back a join call specifying this same unicast address. It can optionally eliminate the random delay in responding to the leave because only one endstation is likely to be responding in this case.

An endstation may optionally support pruning multicast traffic to the source by monitoring the EGMP join and leave calls. If no reception of a packet to a given multicast address from the source is desired according to the EGMP traffic, the endstation can drop the packet without even forwarding it onto the LAN segment. The endstation can use the same algorithm as the switch to determine if packets for that address should be forwarded to to the LAN segment. It must be possible to disable this feature in the endstation when it is attached to a LAN segment with no EGMP-savvy switch. An EGMP-savvy switch can periodically issue a join-all on a LAN segment containing EGMP-ignorant endstations to ensure EGMP-savvy endstations on this LAN segment do not prune back to the endstation.

[6]In expected usage, this case can arise only when another endstation erroneously sends a unicast leave call for another endstation's address.

Normally, a switch just prunes traffic from a given LAN segment in response to traffic load generated by an endstation on a segment sent to an address with no local or remote members. This optimization is of primary interest to endstations such a video servers that send a large amount of multicast traffic and may not know how many members there are to each multicast address at the application level.

4.4.7 Switch `leave` Calls

The switches use the leave-all call to prompt endstations to periodically rejoin groups in which they are still interested, allowing the switches to garbage collect memberships that are no longer of interest.

The interrogator, a distinguished EGMP server switch on the LAN segment, periodically invokes a leave-all call on this segment addressed to the EGMP Ethernet multicast address. This call effectively proposes terminating all memberships on the LAN segment over which it is sent, requiring members to rejoin within the number of microseconds specified in the delay field, the leaveAllDelay. The interrogator then delays for leaveAllDelay microseconds, waiting for join calls.

If no join is received after leaveAllDelay microseconds for a particular address on a given LAN segment in response to a leave-all call, the interrogator invokes a leave call the same as if it received a leave call from an endstation, as described above.

The set of sources for a given multicast address is the union of those specified by the join calls received in response to the leave-all. If this result suggests eliminating one or more sources relative to those currently being delivered, the leave call is retransmitted as described above. Thus, an endstation is given the opportunity to override any reduction in the packet delivery using a join, assuming no more than a single packet loss.

The value of leaveAllDelay should be no more than 1/20 of the leaveAllPeriod used by the interrogator. For example, if the leaveAllPeriod is 20 seconds, the leaveAllDelay should be no greater than 1 second.

Limiting this time period to 1/20 of the leaveAllPeriod means that the delay between a member endstation losing interest in a membership (without sending a leave call, such as by crashing) and the switch stopping the packet forwarding for this address is dominated by the leaveAllPeriod. For example, with a leaveAllPeriod of 20 seconds, an endstation stops receiving packets that were sent to a multicast address on average 10 seconds before the next leave-all call. The leaveAllDelay and the delay of the second address-specific leave call then add a maximum of 2.0 seconds to the time to shut down reception of packets sent to this address.

The second leave call before shutdown is used to ensure that a LAN segment is not incorrectly disconnected from a multicast address or one or more sources as a result of a single packet loss, just as with an endstation-invoked leave operation.

A switch can also send a leave call specifying one or more unicast addresses. In this case, it expects join calls for the designated unicast addresses from endstations. The leave-all-unicast call is interpreted as "memberships for all unicast addresses are expiring"; every endstation on the LAN segment should send a join call specifying its unicast address(es) in response to this leave call. This call is used to quickly learn the location of endstations when a switch first boots.

The unfiltered leave call of a specific unicast address can be used to locate an endstation with this address, thereby avoiding broadcasting the packets to this address. It can also be used to check whether an endstation whose address has been learned supports EGMP (because responding to this leave call indicates it does support EGMP).

4.4.8 Use of Source Filtering for Route Pruning

The join-including or join-excluding calls can be directed at a particular switch by unicasting the call to the switch. In this case, the call removes the path from the switch to the client as part of the specified source(s) multicast tree. This mechanism has potential application in pruning multicast delivery trees in a multi-switch configuration to avoid duplicate delivery. However, this use is for future study.

4.4.9 Comparison to IGMP Source Filtering

The EGMP source filtering is similar to the IGMP version 3 source filtering. However, EGMP uses separate procedure calls for source filtering, allowing it to use lists of multicasts addresses in the join call and leave in the base protocol, unlike IGMP which requires a separate message for each join call. If relatively few memberships use source filtering, the expected case, EGMP results in fewer messages. Moreover, by having the members on a LAN segment agree on the sources by overriding their local exclusions according to the calls by others, the common level of packet traffic on a LAN segment is one leave call and one join-excluding/join-including call per joined multicast group with source filtering.

EGMP also differs from IGMP because a join call in EGMP never reduces the packet delivery, so it is just a performance optimization for other endstations to receive and processing this call. That is, the endstation purely monitors these calls to avoid sending a duplicate join in the case of a leave-all being issues by the switch.

4.4.10 Initial membership Behavior

When an endstation enables its filters for a given multicast address, it issues a join call for that address, typically a join-unfiltered. If the call times out without receiving a response, the client may assume that there is no EGMP-savvy switch on the LAN segment.

When an endstation enables its Ethernet interface for reception, it should send out a join call for the unicast address(es) associated with the interface. This initialization is viewed as an initial join of the endstation address to the network. The join call indicates to the switch that the address is now available on this port.

4.4.11 Pruning to the Endstation

EGMP is defined so that an endstation can perform the same EGMP processing as a non-interrogatory switch to determine whether it needs to forward packets addresses to a particular multicast address to its attached LAN segment. In particular, if there is no member for a given multicast address other than itself based on EGMP packets on this LAN segment, the endstation can drop these packets rather than transmitting them on the LAN segment.

This behavior is possible because EGMP requires that packets for a given address be requested if they are desired. This is true for endstations; it is also true for switches. In particular, a switch requests for each of its LAN segments all packets it needs to receive in order to forward, possibly using the join-all call. To avoid suppressing join calls from endstations, the switch join-all call specifies the exclusion of all addresses that endstations on the LAN segment are members of. It also delays sending the join-all until late in the leave-all period to avoid suppressing join-specific calls. A switch can also just join the specific groups that it needs to forward if those are known, i.e. there are no join-all memberships on its ports and each port is in EGMP mode, as opposed to broadcast mode.

4.5 Timer Values

EGMP uses a number of timer values, as summarized in this section.

leaveDelay - The period of time that a switch waits after receiving a leave-specific call and generating leave-specific calls for group memberships that have not been rejoined on a LAN segment. It also waits leaveDelay microseconds after (re)issuing a leave-specific call before stopping the forwarding of packets, assuming no subsequent join for that address is received. The leaveDelay value is also the time used by the switch between retransmitting such a call and having the restricted source filtering take place, assuming there are no calls received that further modify the source filtering.

leaveAllDelay - The period of time that a switch waits between invoking a leave-all call and generating leave-specific calls for group memberships that have not been rejoined on a LAN segment.

leaveAllPeriod - The period of time that the interrogator waits between issuing a leave-all call on a LAN segment to prompt endstations to rejoin groups. It is also the time interval between which endstations and switches retry to prune reception of packets after an earlier request was refused or overridden.

The leaveDelay value is effectively the time period used when reducing the packet forwarding to a LAN segment when the reduction is expected to effect a single or small number of endstations. The leaveAllDelay is the value used when the action is affecting all the endstations on the LAN segment.

As mentioned earlier, the value of leaveDelay is recommended to be at least 10 times the maximum packet transmission time, *e.g.,* 1.2 milliseconds for 100 Mb Ethernet and 12 milliseconds for 10 Mb Ethernet. The switch should increase this value if the LAN segment is shared, under heavy traffic, or the leave call specifies a large number of addresses.

The value of leaveAllDelay should be no more than 1/20 of the time period between leave-all's issued by the interrogator. For example, if the leave-all period is 20 seconds, the leaveAllDelay should be no greater than 1 second.

The interrogator may use an adaptive algorithm to compute and revise the leaveAllDelay it uses. For example, it could use a shorter leave-all period as the multicast traffic increases so that the leave-all overhead remains a small percentage of overall multicast traffic, and also shorten the leaveAllDelay value dynamically until it appears too short. It is too short when it either receives no membership calls for some address within leaveAllDelay in response to the leave-all call and does receive a join call later, or else receives multiple join calls (indicating the leaveAllDelay is too short to have the randomized delay suppress duplicates). Using this adaptive approach, EGMP can provide the lowest leave latency that is efficient for the endstations on the LAN segment that cuts off extraneous multicast traffic as quickly as possible. Because this adaptivity appears to be an unnecessary complication at current levels of multicast traffic, its implementation is considered optional at this time. (The clients automatically adapt to the server behavior by using the delay values used by the servers.)

The leaveAllPeriod is chosen to tradeoff the time to garbage collect group memberships versus the overhead on the LAN segments and endstations of effectively requerying the membership information. The EGMP leaveAllPeriod is 3 minutes. Al switches should use this value to keep the choice of interrogator stable over time, except for failures. (A switch can optionally invoke a leave-specific on a very high-demand group more frequently if so desired.)

A switch should add an extra processing time to the leaveDelay and leaveAllDelay times that it uses internally relative to those it advertises to the clients so that a client join call that is randomly delayed by the maximum time (according to the values of leaveDelay or leaveAllDelay) is received before the switch actually times out the membership, allowing for expected queuing and processing delays.

4.6 Per-LAN Segment EGMP Interrogator Election

EGMP tries to operate with just one EGMP interrogator per LAN segment. This is accomplished using an election mechanism as follows. Each switch (or bridge) functions initially as an interrogator on each of its ports. However, if a switch sees a leave-all call with a non-zero delay from an Ethernet address that is lower than its own Ethernet address, it stops acting as an interrogator on that LAN segment until it has not seen a leave-all call again for at least two leave-all intervals, at which point it resumes again.

As a suggested implementation, the switch sets a timer for twice the leaveAllPeriod when it receives a leave-all call on a port from a lower-addressed source and stops acting as an interrogator on that port. If it receives a subsequent leave-all call on this port from the interrogator during this time interval, it records that fact in a per-port flag. When the timer expires, the server timer routine checks whether leave-all calls have been seen on this port during the time-out period. If yes, the timer is reset to twice the leave-all call interval and the switch continues as a non-interrogator. Otherwise, the switch reverts back to acting as an interrogator, assuming the previously selected interrogator has failed.

This procedure is similar to that used in IGMP version 2, minimizing the overhead on a shared LAN segment containing many switches. It also eliminates the need to prevent "convoying" of leave-all calls, as can arise when multiple switches are serving as interrogators on the same LAN segment. Packet convoys result when the switches unintentionally self-synchronize over time so that the set of leave-all calls are transmitted one right after the other over the LAN segment.

4.7 Restricted Multicast

In some environments, the network administrator may wish to preclude unauthorized endstations from joining particular groups. EGMP assumes that these restrictions are specified to the switch using a separate management mechanism. The switch can then refuse join calls from unauthorized endstation interfaces and LAN segments, returning a negative response.

As currently specified, EGMP does not support authentication. However, it would be straightforward to support one or more of the standard ONC RPC authentication mechanisms. This authentication support in conjunction with separately specified memberships access controls allows EGMP to support restricted access multicast. In this case, a membership call returns with an error indication rather than the packet delivery simply not working. This return indication associated with membership allows the client to distinguish between a switch granting the membership, refusing the membership and not responding.

5 Supporting IP Routers and Similar Devices

An IP router needs to receive multicast packets sent to any addresses in the Ethernet multicast address range designated for IP multicast. An IP router uses EGMP to receive this range of addresses as follows.

The IP router invokes a join-all call, indicating that it wants to receive all multicast packets. Subsequently, when the router receives a packet addressed to a multicast address outside the range that it is interested in, it sends a leave call to the switch, specifying this multicast address. The switch performs the standard leave processing specified above, ultimately blocking further transmission of packets to this multicast address over this LAN segment if there are no join calls received for this address from the LAN segment. (The switch can record this "prune" internally either as an explicit exception to the join-all call, or by removing this LAN segment from the list of ports for the multicast address.)

After pruning one or more addresses, the router responds to subsequent leave-all calls with a join-all listing the pruned addresses in the descriptor. Thus, if packets sent to the unwanted multicast address continue to arrive, the pruning is retried every leaveAllPeriod microseconds. The router should not re-request a leave when unwanted multicast packets continue to arrive to avoid extra traffic in the case that the switch does not support this filtering or there are other interested parties on the same LAN segment. In this vein, the leave call is just an optimization to improve performance.

This "receive-all-and-prune" approach can be used by routers for other protocol architectures that support multicast. It can also be used by network sniffers that are monitoring multicast traffic. However, note that an endstation or router that does a join-all call must operate in multicast promiscuous mode to detect the full range of multicast packets being forwarded on its LAN segment. Moreover, the total amount of multicast traffic that is forwarded in response to a join-all call may exceed the capacity of the LAN segment. In the preferred configuration, an endstation or router using a join-all call is connected to a switch by a LAN segment with no other endstations, switches or routers on it, operates in multicast promiscuous mode, and uses the leave call to avoid overloading its LAN segment. We suggest that switches provide the management option to ignore join-all calls on some ports, so the network administrator can prevent random endstations from using this facility.

6 Inter-Switch Operation

EGMP also serves as an inter-switch multicast membership protocol for multi-switch configurations. The calls sent between switches are the same as described earlier except they are addressed to a separate EGMP inter-switch multicast address. They are also processed by the switch slightly differently than an endstation. Essentially, a switch serves as an EGMP proxy for the endstations that connect to it, joining groups to receive multicast packets from other switches and forwarding these packets to its attached endstations as indicated by their memberships.

Each port of a switch is set to operate on one of two modes for its inter-switch operation. In the first mode, the *broadcast-and-prune* mode [7], a switch uses a join-all call to join to all multicast groups on an attached LAN segment. It then prunes reception of packets addressed to multicast addresses for which it has no memberships, the same as described earlier for routers. In the second mode, the *specific membership* mode, a switch specifies precisely the groups for which it has membership requests from its attached endstations. If a switch receives a join-all on one of its ports, whether from another switch or a router, it must issue a join-all to all other switches from which it receives packets.

The switches are assumed to be running a spanning tree algorithm or a distributed routing algorithm so that they avoid packet loops and duplicate delivery. This document assumes the use of the standard spanning tree algorithm[15] for this purpose. The use of other routing mechanisms is for future study. However, in the following discussion, we use the term "ports leading away from the source" to indicate the set of ports to which the switch would normally forward a broadcast packet from the given source, suggestive of the greater generality for inter-switch operation that we expect to be developed in the future. EGMP can be used with multiple spanning trees, one per virtual LAN, using the virtual LAN-specific form of EGMP, as described in Section 6.7.

Unless otherwise stated as an endstation call, all calls in the following subsections are inter-switch calls using the EGMP inter-switch address.

6.1 The `join` Call

For each port in broadcast-and-prune mode that is actively receiving for the spanning tree, a switch issues a join-all with a delay value of 0, indicating unbounded membership. The join-all call lists the multicast addresses that have been pruned by leave calls.

It reissues the join-all call following the reception of a leave-all call.

This join-all call causes switches attached to the LAN segments of this switch to forward all multicast packets to the joining switch unless they are explicitly pruned by the exclusion list in the join-all call.

For each port in specific membership mode, the switch issues a join-specific call listing each address that the switch needs to forward. If a switch receives and accepts a join-all on one of its ports, it needs to issue a similar join-all at the endstation and switch levels for each port that it can receive and forward packets, within the constraints of the appropriate spanning tree.

6.2 `join` Call Handling

When a switch receives a join-all call, it records the need to forward all multicast packets to this port and sends back a response. It also flags the port as connecting to a switch. If this switch is the interrogator switch for this port, it (re)invokes the join call to ensure all switches received the membership, the same as the endstation protocol except for using the inter-switch multicast address.

The interrogator switch for a port can serve as the interrogator for the inter-switch expiration of memberships on that port because it was logically elected by the same algorithm.

A join-specific call on a port P is handled the same as an endstation join call except the membership is flagged as at the switch level.

When a switch receives a join-specific call on a port P on which it previously received a join-all call, the switch ensures it is forwarding packets to this address out P by removing any record of a leave for this address at P and creating a membership record for this address. It also issues a join call for this address on each port whose packets to this address would now be forwarded to port P (to make sure any pruning of this address is undone).

If the switch has not received a join-all call on port P, it creates a record for this new membership on this port. It also ensures that it is a member of this group on all ports from which it would forward packets to this address to port P. In the normal case, this forwarding information is provided by the spanning tree algorithm. In effect, the join call is the means for unpruning the multicast distribution for an address when a new member appears. In the broadcast-and-prune model, a member effectively listens to an address by unpruning the tree forward to itself. (The unpruned address is removed from the exclusion list of the join-all call issued in the next leave-all period.)

The broadcast-and-prune approach is preferable when members are joining multicast groups that are largely inactive, so no pruning is required. The specific membership model seems preferred if many of the addresses are active but without members. In this case, state is only created for the groups with members. With broadcast-and-prune in this case, the switches would store a large amount of pruning state. However, the specific membership model seems harder to make robust because the expected failure mode leads to packet loss, not just excessive broadcasting, as with the broadcast-and-prune approach.

6.3 The `leave` Call

A switch invokes a leave call on a LAN segment when it is receiving packets for a multicast address that it has no interest in receiving. The leave call is sent with a delay value of 0, just as with endstations, to flag this call as a "client" request rather than an interrogator call.

If the leave fails to stop the packets from arriving, the switch does not reinvoke the call, the same as with the endstation-to-switch leave protocol. Instead, in subsequent leave-all periods, it requests memberships that do not include the undesired addresses. As with the endstation protocol, this approach provides low leave latency in the common case and yet avoids extra (futile) packet overhead when there are other stations on the LAN segment that need to receive the traffic that this switch does not want to receive.

6.4 The `leave` Call Handling

When a switch receives a leave call on the EGMP inter-switch multicast address, it removes the corresponding inter-switch join record if any and otherwise ignores the request if there is an endstation membership for this address on port P. It otherwise uses the same leave procedures and timing as described for the endstation-to-switch protocol, but operating as both switch and endstation and using the EGMP inter-switch protocol.

In particular, as an endstation, if it has a membership in one or more groups specified in the leave call, or wants to receive from sources that are to be filtered according to the leave call, the switch invokes a join call overriding these aspects of the leave call. As a switch, if it is an interrogator, it delays and then sends a leave call for these groups if it does not receive a join call for them before the timeout.

A switch should not respond to an interrogator leave as an interrogator to avoid livelock between multiple switches acting as interrogators. A switch can distinguish a leave sent as a client request and a leave call sent by an interrogator by whether the delay field is non-zero or not. A zero value indicates that it was sent by a "client" switch, the same as for the endstation protocol.

6.5 Inter-switch Packet Forwarding

When a switch receives a data packet addressed to a multicast address, it forwards the packet on each port that is part of the spanning tree and has a membership for this packet, excluding the port on which the packet was received. The membership can either be a specific membership for this address or else a join-all with no leave record (prune) for this particular address. In the broadcast and prune mode, the initial packets to an address are broadcast to all the switches until pruning takes place.

When a switch receives a packet for some address M and does not need to forward it, it can send a leave call specifying that address to the port on which the packet was received provided that it has not sent such a leave call in the last leave-all period. It then adds this source address to the list of pruned addresses for that destination address and this port.

6.6 Broadcast-and-Prune versus Specific membership

A switch can monitor the packet overhead it is incurring on a port in the broadcast-and-prune mode and switch to specific memberships for that port if the specific membership mode is less expensive. With broadcast-and-prune, the switch and port incur the overhead of receiving unwanted multicast packets, leave calls to prune the traffic, and join calls to unprune the traffic once it is requested. If there is significant traffic on many multicast addresses with sparse members, there is considerable overhead for pruning the traffic in the broadcast-and-prune approach. (There is also the overhead of unpruning when a particular group is joined, and pruning it again when it is left, but that is comparable to the explicit membership overheads.) However, if there is relatively little traffic on multicast addresses (so no pruning is needed) yet there is a high rate of joining and leaving groups, the broadcast-and-prune approach can be less expensive than using specific memberships.

A switch can only change to using explicit memberships on a port if it is not receiving a join-all membership on any other port (for otherwise it has to receive and forward all multicast). As a consequence, changing from broadcast-and-prune to specific membership would generally occur at leaf switches (of the spanning tree) first and possibly propagate to the intermediate switches from there. That is, the leaf switch learns the specific multicast addresses of interest to its LAN segments that only connect to endstations, assuming there are no join-all memberships on these segments. It then changes to specific membership mode on the link to another switch. Once each leaf switch connected to some interior switch changes to specific membership mode, the interior switch should have a single actively receiving link to a next level switch, and it can change this link into specific membership mode as well. Thus, the whole spanning tree of switches can change to specific membership mode from the leaves inwards if this mode is supported and favored by all the switches according to the traffic conditions.

The implementation of the specific membership mode is optional. However, a switch must be able to operate in broadcast-and-prune mode, including switching to this mode from specific membership mode so that it can interoperate with other switches using broadcast-and-prune.

6.7 Virtual LANs and Inter-switch Operation

A virtual LAN primarily defines a broadcast domain. A multicast on a virtual LAN is sent to the subset of those endstations in the virtual LAN's broadcast domain that join the specified multicast address. Given that virtual LANs can span two or more switches, EGMP needs to support the forwarding of multicast traffic between switches that is consistent with (distributed) virtual LAN semantics. In particular, a packet should only be delivered to ports or endstations that are in a common virtual LAN with the source specified in the packet. With the evolving state of distributed virtual LAN management protocols at the time of writing, this document does not provide one fixed solution. However, EGMP can support distributed virtual LANs in two ways.

In the first approach, inter-switch EGMP calls can be made virtual LAN-specific by the switch using a source address in the call packet that is coupled to the virtual LAN of the membership. A switch can then determine the virtual LAN associated with a call by determining the virtual LAN associated with the source address of the call packet. Mapping source addresses to virtual LAN seems necessary unless some encapsulation and tagging scheme is used, like the proposed IEEE 802.10-based approach. In the latter case, the packet could rely on the encapsulation scheme to specify the virtual LAN associated with the EGMP calls.

This approach is preferred because it fits the model of treating each virtual LAN as an independent broadcast domain and using EGMP to selectively multicast within each such domain. In particular, there appears to be an EGMP server per virtual LAN on each switch.

As an alternative approach, a switch can use virtual LAN-independent memberships (as described by this document to this point) and simply discard traffic from one virtual LAN whose only members local to this switch are in a different virtual LAN. If the switch detects that it is receiving an excessive amount of traffic that it is discarding for this reason, it uses source filtering to request that the sending switch eliminate the source(s) providing the traffic it has to discard. This call may be unicast to the sending switch. (It is assumed that one or a small number of sources account for this traffic load that is to be pruned.) The traffic is thereby pruned to fit the virtual LAN configuration rather than specifying the virtual LAN(s) associated with each membership.

With these two options available, there appears to be no reason to explicitly extend EGMP to support distributed virtual LANs, no matter how the distributed virtual LAN management protocols evolve. However, this is an area for further study.

6.8 Switch Use of Unicast Queries

A switch invokes a leave call specifying a unicast address to the inter-switch EGMP multicast address to locate one or more specific unicast addresses on switches attached to its ports. (Logically, this call just notifies the other switches that the sending switch is planning to stop forwarding packets for the given destination unless a unicast membership for this address is created.)

Normally, a switch first invokes a leave call for the unicast addresses using the endstation EGMP multicast address to determine whether these addresses correspond to directly connected endstations. If this leave call fails to generate a response, it then uses the inter-switch address to query at the switch level.

A switch receiving a leave call for a unicast address sent to the EGMP inter-switch multicast address checks whether this unicast address is local to this switch. If so, it sends a join call on the EGMP inter-switch multicast address specifying this unicast address to the port on which the leave call was received, provided that this address is not on the same port as the leave call was received. (If the latter, the leave call is ignored because the endstation on that port should have responded to an earlier leave call sent to the endstation EGMP multicast address.)

If the address is not known as local to the switch, the switch creates a record for this unicast address, marking it as requested-leave. The switch then broadcasts the leave call on all branches of the spanning tree except for the port on which it was received. It may also (concurrently) invoke a leave call using the EGMP endstation multicast address to check whether the unicast address is that of an endstation directly connected to this switch.

When the switch receives a join call on the inter-switch EGMP multicast address, it records the port on which the join call was received as corresponding to this unicast address, or takes some error reporting action if this mapping is inconsistent with its configuration[7]. If this address record is marked as being requested by another switch, this switch sends a join for this unicast address to the requesting switch.

A switch receiving a leave call for a unicast address on the endstation EGMP multicast address responds as described in Section 4.4.5.

This unicast leave call mechanism allows a switch to locate the endstation corresponding to a particular unicast address, even in a multi-switch configuration, before forwarding packets to this address.

In summary of the inter-switch use of EGMP, a switch is able to interoperate with other switches to selectively forward multicast packets using EGMP as described above and the standard spanning tree protocol to avoid packet loops. It may also be feasible to use EGMP in conjunction with a separate proprietary routing system between switches that understands this routing system, providing that this mechanism allows a switch to distinguish which ports are towards a given address and which are away from this address. Source filtering support at the inter-switch levels seems like an important mechanism for pruning multicast trees in this case, judging by the experience with DVMRP. However, this is an area for further study.

7 Implementation

EGMP has both a client (or endstation) and a server (or switch) implementation.

7.1 Client EGMP

Client EGMP is implemented in the network driver for each interface of an endstation. The network driver is informed by the higher layer to start receiving a particular multicast address, causing it to create a record for that multicast address and send out an initial join call on this membership. Similarly, the driver sends out a leave call when the higher-level software instructs it to stop receiving on a particular multicast address. Rather than implement all of this directly in the Ethernet driver, the driver can be modified just to call functions in the EGMP module interface on these actions which make the appropriate remote procedure calls to add and delete memberships. Implementing EGMP in the device-independent portion of the Ethernet driver seems appropriate, as is conventionally done with ARP.

A client must implement all of the server RPC calls as described in the protocol description. However, the processing of these calls is relatively straight-forward because the actions are in terms of the key client data structures and calls. For example, receiving a leave call may cause the client to search its membership data structures and possibly generate a join call. However, this is something it must implement in any case.

A client need not implement all of EGMP to be EGMP-compliant. Pruning to the endstation is optional. Support for source filtering is also optional, except the client must receive and respond to leave calls that would otherwise restrict the packets the client is to receive. The client simply responds with a join that requests receiving for that address from all sources in response to any attempt to apply source filtering to that multicast address.

[7]For example, a switch might be user-configured with this unicast address on another specific port.

7.2 Server EGMP

The server EGMP is implemented in the switch or associated agent that is managing the switch. The EGMP agent must be able to send and receive packets on particular ports of the switch as well as control the switch forwarding of multicast packets according to the memberships it maintains. Server EGMP also includes inter-switch EGMP.

A server should support three modes for each port:

Normal - forward no unrequested multicast to this port and obey all EGMP packets from this port. This mode assumes all endstations connected to the port are EGMP-savvy.

broadcast - forward all multicast to this port, ignoring EGMP packets from this port. This mode allows EGMP-ignorant endstations to be connected even if there are other EGMP-savvy endstations on the same LAN segment. The switch needs to send a join-all call periodically on this LAN segment on behalf of the EGMP-ignorant endstations to ensure that other EGMP endstations on the LAN segment do not prune their transmissions back to their interface. The switch still only forwards those packets for which it has memberships elsewhere, independent of this join-all call.

ignore-join-all - The same as the Normal mode except that join-all calls are ignored. This mode simply restricts endstations on the port from using the join-all call to receive all multicasts. For example, if a port is connected to a hub shared by several clients, usage of a join-all call by one of the client machines may be inappropriate.

A server need not implement all of EGMP to be EGMP-compliant. In particular, pruning to the endstation and source filtering are optional.

An EGMP server must also implement the the EGMP client portion for inter-switch operation.

7.3 Reference Implementation

A reference implementation of client EGMP is 900 lines of C++. Server EGMP is approximately 1500 lines of C++. *These numbers could vary somewhat based on differences in the implementation environment.*

8 Concluding Remarks

EGMP is a protocol in the spirit of IGMP that allows the switches of an extended Ethernet to determine the LAN segment(s) to which packets addressed to a given address should be delivered, multicast or unicast. EGMP allows switched or bridged Ethernet to avoid the broadcasting of multicast packets and packets addressed to unknown unicast addresses, thereby eliminating a key impediment to scaling for extended Ethernets.

With the extensions for source filtering, EGMP supports control of packet delivery based on *all* fields defined at the Ethernet level except for protocol type. We argue that protocol-specific subscriptions are not necessary because in practice, each protocol uses a specific subset of the multicast addresses that does not overlap with that used by others. Therefore, the protocol type is implicit in the multicast address and protocol-specific membership would only be useful to guard against random garbage traffic, which is infrequent. One might also argue that the length of the packet is also available at the MAC level, but there is no clear use for filtering based on packet length. Therefore, there seems little reason to extend EGMP further for MAC-level control.

EGMP is designed in part to support IP multicast on extended LANs. However, it can also be used for other protocols such as Appletalk, IPX, etc.

EGMP is specified for Ethernet. It can be directly used with other network technologies that use Ethernet address formats, such as FDDI. Its design could be readily adapted to other switched network technology with similar broadcast/multicast issues. The experience with ARP, which was designed for other technologies but not really used for same, suggests specializing EGMP for Ethernet to avoid ARP-like "type of address" fields. We leave it to those working with separate MAC protocols and other network technologies to adapt EGMP if needed and so desired. Hopefully, the design of EGMP as an RPC-generated protocol based on GMP will facilitate this adaptation.

Acknowledgment

Thanks to Jonathan Stone for his review of any earlier draft of this document and his idea of using the source address to identify the virtual LAN in inter-switch use of EGMP with distributed virtual LANs.

References

[1] F. Backes, "Transparent bridges for Interconnection of IEEE 802 LANs", IEEE Network, 2(1), 5-9, January 1988.

[2] T. Bradley and C. Brown, Inverse Address Resolution Protocol, RFC 1293, Jan 92

[3] B. Cain, A. Thyagrajan, S. Deering, Internet Group Management Protocol, Version 3, draft RFC, private communication.

[4] D.R. Cheriton, Recursive Structuring of an RPC Protocol Architecture, Proceedings of SIGCOMM'88, ACM, Stanford, CA 1988.

[5] S. Deering and R. Sharma, Planned and Possible Changes to IGMP, Version 2, slides from IDMR Working Group Meeting, March 1994.

[6] S. Deering, "Host Extensions for IP Multicasting", RFC 1112, Stanford, Aug. 1989.

[7] S. Deering, "Multicast Routing in a Datagram Internetwork", Ph.D. thesis, Stanford University, available as STAN-CS-92-1415, December, 1991.

[8] J. Hart, "Extending the IEEE 802.1 MAC Bridge Standard to Remote Bridges", IEEE Network, 2(1), 10-25, January 1988.

[9] IEEE P802.1d/D0, Draft Stanford for Traffic Class and Dynamic Multicast Filtering Services in Bridged Local Area Networks (Draft Supplement to 802.1D), September 26, 1995.

[10] M. Laubach, Classical IP and ARP over ATM, RFC 1577, Jan. 1994

[11] J. Mogul and J. Postel, "Internet Standard Subnetting Procedure", RFC 950, Stanford University and USC/ISI, Aug. 1985.

[12] D. Plummer, An Ethernet Address Resolution Protocol or, Converting Network Protocol Addresses to 48-bit Ethernet Addresses for Transmission on Ethernet Hardware, RFC 826, Symbolics, Nov. 1982.

[13] J. Postel, "Multi-LAN Address Resolution", RFC 925, USC/ISI, Oct. 1984.

[14] C.Smoot and J. Quarterman, "Using ARP to implement transparent subnet gateways", RFC 1027, Oct. 1987.

[15] ISO/IEC 10038 [ANSI/IEEE Std 802.1D 1993 Edition], Information Technology — Telecommunications and information exchange between systems — Local are networks — Media access control (MAC) bridges.

[16] Sun Microsystems Inc, "XDR: External Data Representation Standard", RFC 1014, Sun Microsystems Inc, June 1987.

[17] Sun Microsystems Inc, "RPC: Remote Procedure Call Protocol Specification Version 2", RFC 1057, Sun Microsystems Inc, June 1988.

A RPC Implementation

EGMP requires a minimal ONC RPC run-time implementation.

The client RPC run-time is expected to retransmit a non-datagram call until receives a reply or has retransmitted 5 times, with a retransmit interval of callRetransmitTime, where callRetransmitTime is the maxDelay value used by the interrogator for the last leave-all call, or else 20 milliseconds if the former is not known. The client RPC run-time should also cause the *xid* field of each call packet to be one greater (or more if necessary) than the previous call from this client. A response should be matched to the call using the *xid* field. Duplicate responses can be discarded by the call record being marked as discarded (or just the local variable storing the last *xid* value used being incremented).

The server RPC run-time can be quite simple, recognizing that the EGMP calls are all idempotent. That is, there is no need to do duplicate suppression on calls. If a duplicate call request packet is received, the call can just be performed again to regenerate the response, eliminating any need to save the response to a call for retransmission or the state and code to detect duplicates. There is also no need to enforce the ordering of calls based on the sequence numbers, because calls are unlikely to be reordered because they traverse a single LAN segment in general. Also, reordering of calls does not have a negative effect over the expected time intervals for MAC-level communication. That is, a leave call that was issued before a join call for the same address, but is reordered to occur later, causes the member to rejoin either in response to receiving the delayed leave call or in response to the interrogator switch retransmitting the leave call in response to receiving the call. A join call that is reordered to occur later than a leave call for the same address that was issued later may result in an unwanted membership. However, this membership would be recollected in the next leave-all period so the inconsistency is short-lived.