Network Working Group
Request for Comments: 2443
Category: Experimental

J. Luciani
Bay Networks
A. Gallo
IBM

November 1998

A Distributed MARS Service Using SCSP

Status of this Memo

Copyright Notice

Abstract

This document describes a method for distributing a MARS service within a LIS[1]. This method uses the Server Cache Synchronization Protocol (SCSP)[2] to synchronize the MARS Server databases within a LIS. When SCSP is used to synchronize the caches of MARS Servers in a LIS, the LIS defines the boundary of an SCSP Server Group (SG).

1. Introduction

The keywords MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD, SHOULD NOT, RECOMMENDED, MAY, and OPTIONAL, when they appear in this document, are to be interpreted as described in [5].

The MARS is an extended analog of the ATMARP Server introduced in [4]. It provides the necessary connection and addressing services required by layer 3 multicast services over ATM. There are three basic elements to the MARS model. First, the MARS Server which manages and distributes layer 3 group membership information to the LIS. Second, MARS Clients which register with and query a single MARS Server for layer 3 multicast information. Third, MCS Clients which register with a single MARS Server and provide layer 3 multicast forwarding services for a LIS.

Both MARS Clients and MCS Clients explicitly register with the MARS Server before exchanging layer 3 multicast information. During the registration process MARS Clients are place on the Cluster Control VC

(CCVC) and MCS Clients are placed on the Server Control VC (SCVC). Both the CCVC and SCVC are then used to propagate layer 3 multicast updates to the clients which make up a LIS. During the registration process MARS Clients are also assigned a unique Cluster Member ID (CMI) which is used to identify reflected packets in the presence of MCS Clients.

In the Distributed MARS Model there MAY be multiple MARS Servers in a given LIS, and since any MARS Server within the LIS MUST be able to provide layer 3 multicast information about any multicast group within the LIS, there MUST be a method by which to synchronize multicast information across all MARS Servers within the LIS.

The Server Cache Synchronization Protocol (SCSP) solves the generalized server synchronization/cache-replication problem for distributed databases, and thus SCSP MAY be applied to the MARS Server database synchronization problem within a LIS. When SCSP is used to synchronize the caches of MARS Servers in a LIS, the LIS defines the boundary of and SCSP Server Group (SG).

SCSP is defined in two parts: the protocol independent part and the client/server protocol specific part. The protocol independent part is specified in [2] whereas this document will specify the client/server protocol specific part where the MARS Server is the client/server protocol.

2. Overview

All MARS Servers belonging to a LIS are said to belong to a Server Group (SG). A SG is identified by, not surprisingly, its SGID which is contained in a field in all SCSP packets. All SCSP packets contain a Protocol ID (PID) field as well. This PID field is set to 0x0003 to signify that SCSP is synchronizing MARS Server databases as opposed to synchronizing some other protocol's databases. (see Section B.2.0.1 of [2] for more details). In general, PIDs for SCSP will be assigned by IANA upon request given that a client/server protocol specific specification has been written. In the case of MARS Servers, the client/server protocol specific specification was written at the same time as SCSP, and thus a PID=0x0003 was assigned in [2].

SCSP places no topological requirements upon a MARS Server SG. Obviously, however, the resultant graph of MARS Servers must span the set of MARS Servers being synchronized. For more information about the client/server protocol independent part of SCSP, the reader is encouraged to see [2].

When a SG is using SCSP for synchronization, a MARS Client or MCS Client will register with only one MARS Server although it is allowed to choose any MARS Server in the SG for this registration. At registration time the MARS Client or MCS Client will be added to that MARS Servers respective CCVC or SCVC. Also, MARS Clients will be issued a unique CMI for the entire LIS. This document assumes at a minimum each MARS Server in the SG will be configured with a unique range of CMIs to assign to clients registering with that MARS Server. Use of some external means for allocating CMIs to MARS Servers in a SG is possible but beyond the scope of this document.

When a MARS Client or MCS Client successfully registers with a MARS Server in the SG that MARS Server will propagate the registration information to its peer MARS Servers. The same propagation will occur for any subsequent group membership information learned from the clients. The peer MARS Server will then update its group membership database and propagate the information out its own CCVC or SCVC if needed.

In the case of a MARS Server failure all peer MARS Servers in the SG MUST flush the client/group membership information learned from the failed MARS Server. The clients belonging to the failed MARS Servers CCVC and SCVC will migrate to the next available MARS Server as specified in Section 5.3 of [1]. When a client detects a failure of its MARS, it steps to the next backup MARS Server and attempts to register with the server. If the registration is successful the client will re-join all of its previous group membership information. If the registration fails, the process repeats until a functional MARS Server is found.

Determining the operational state of a MARS Servers in a SG requires that each MARS Server send out an "alive" or "heartbeat" message similar to the MARS Redirect message sent out on the CCVC or SCVC for MARS Clients. However, this message will only be sent to MARS Servers in the SG and is from here on defined as the MARS Server Redirect Entry.

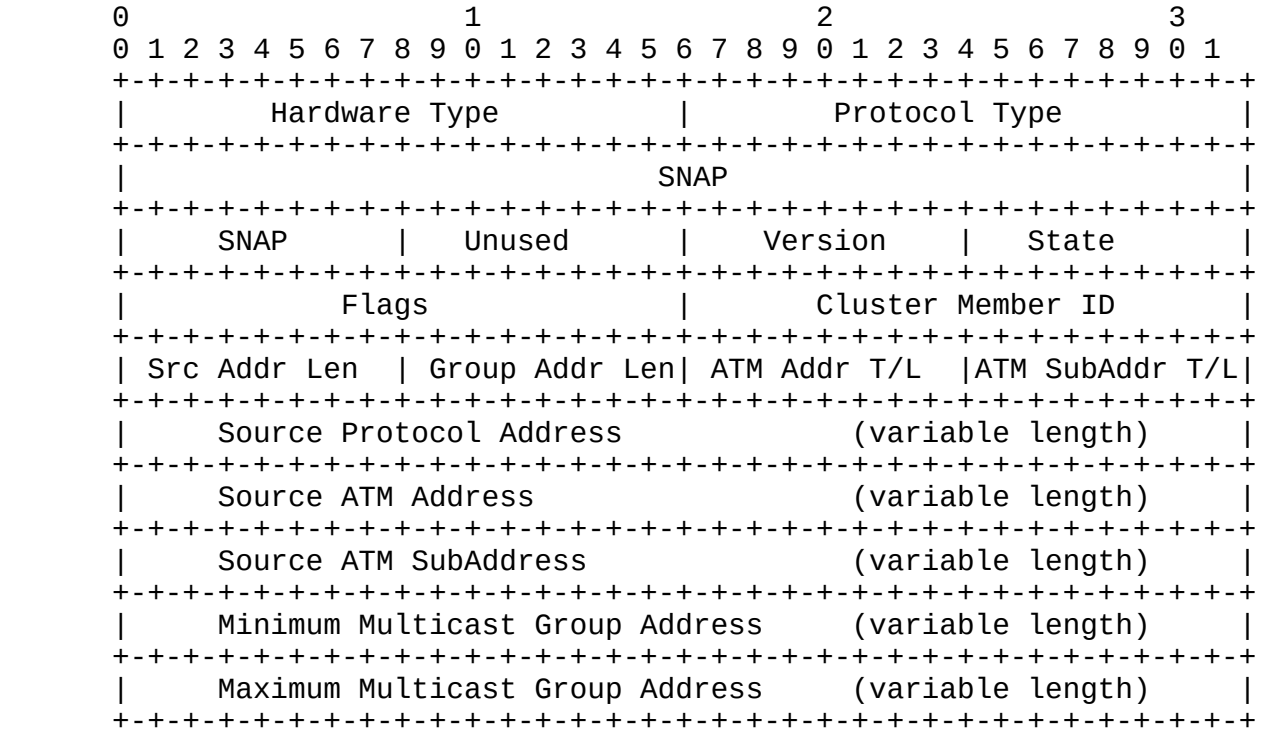
In order to detect that a MARS Server failure has occurred each server MUST update it’s MARS Server Redirect Entry state at least every 2 minutes, it is RECOMMENDED that it is updated every 1 minute. Failure to receive two consecutive MARS Server Redirect Entry updates from a given MARS Server in the SG will cause all membership information learned from this server to be flushed. The MARS Server Redirect Entry state is also used to create the MARS_REDIRECT_MAP messages sent out on CCVC for each MARS Server in the SG. The ordering of each server learned will be based on the MARS Servers SCSP Sender ID. The ordering of the MARS_REDIRECT_MAP will first

contain the list of MARS Servers learned via MARS Server Redirect Entry updates in ascending order based on the SCSP Sender ID, followed by any externally configured or learned backup MARS Servers.

In the case of a MARS Client or MCS Client failure where the client is unexpectedly removed from the CCVC or SCVC the MARS Server MUST notify its peer SG members via a proxy deregister for that client. Upon receiving a proxy deregister request from a peer SG member all membership information for the deregistering client MUST be removed. Any Clients sending multicast data to the failed client should also receive an unexpected removal of this client which will intern cause the sending client to revalidate the multicast groups current membership as outlined in Section 5.1.5.1 of [1].

3. Format of the CSA Record MARS Specific Part

CSA Records in SCSP contain a "Client/Server Protocol Specific Part" which contains the non-protocol independent information for a given server's cache entry.



Hardware Type
Defines the type of "link layer" addresses being carried. This value is the ATM Forum 'address family number' specified in [3] as 15 decimal (0x000F). This is the mar$afn field defined in [1].

Protocol Type
This field is the protocol type number for the protocol using MARS from [3]. (IPv4 is 0x0800). This is the mar$pro.type field from [1].

Protocol SNAP
This field is the optional protocol SNAP extension to protocol type. This is the mar$pro.snap field from [1].

Version Number

0	MARS Specific part of the CSA record.
0x01	Reserved for NHRP.
0x02 - 0xEF	Reserved for future use by the IETF.
0xF0 - 0xFE	Allocated for use by the ATM Forum.
0xFF	Experimental/Local use.

Version Number for this document MUST be set to 0x00.

State
1 MARS Server Redirect Entry.
2 MCS Serve/Register request.
3 MARS Client Join/Register request.
4 MARS Client Leave/Deregister request.
5 MCS Unserve/Deregister request.

All other State values should cause the CSA to be discarded.

Flags
The flags field is used to contain several flags and is similar to the mar$flags field from [1].

mar$flags
Bit 15 - mar$flags.layer3grp
Bit 13 - mar$flags.register
Bit 0-7 - mar$flags.sequence

All remaining bits are reserved and MUST be zero. The mars$flags.sequence field is of local significance only to the Local Server (LS).

Cluster Member CMI
This field contains the CMI which uniquely identifies each endpoint within a LIS. This is the mar$cmi field from [1].

Src Addr Len
This field contains the length of the Source Protocol Address field. For IPv4, the value is 4 if an address is specified. A null (non-existent) address MUST be coded as zero length, and no space allocated for it in the message body. This is the mar$spln field from [1].

Group Addr Len
This field contains the length of the Group Protocol Address field. For IPv4, the value is 4 if an address is specified. A null (non-existent) address MUST be coded as zero length, and no space allocated for it in the message body. This is the mar$tpln field from [1].

ATM Addr T/L
This field contains the type and length of the Source ATM Address field. The type and length encoding is described in Section 5.1.2 of [1].

ATM SubAddr T/L
This field contains the type and length of the Source ATM SubAddress field. The type and length encoding is described in Section 5.1.2 of [1].

Source Protocol Address
This is the internetwork address for the source of an address binding in a MARS server cache entry. If null, no storage will be allocated. This is the mar$spa field from [1].

Source ATM Address
This is the Source’s ATM address of an address binding in a MARS server cache entry. The address, if specified, is E.164 or ATM Forum NSAPA. This is the mar$sha field from [1].

Source ATM SubAddress
This is the Source’s ATM subaddress of an address binding in a MARS server cache entry. The subaddress, if specified, is an ATM Forum NSAPA. If null, no storage will be allocated. This is the mar$ssa field from [1].

Minimum Multicast Group Address
This is the internetwork address of the lower bound on the range of multicast group addresses for the address binding in a MARS server cache entry. If null, no storage will be allocated. This is the mar$min.N field from [1].

Maximum Multicast Group Address
This is the internetwork address of the upper bound on the range of multicast group addresses for the address binding in a MARS server cache entry. If null, no storage will be allocated. This is the mar$max.N field from [1].

4. Values for SCSP Protocol Independent Part

The following sections give values for fields of the SCSP Protocol Independent Part of the various SCSP messages.

4.1 Values for the SCSP "Mandatory Common Part"

Protocol ID = 0x0003
Sender ID Len = 0x04
Recvr ID Len = 0x04

See Section B.2.0.1 of [2] for a detailed description of these fields.

4.2 Values for the SCSP "CSAS Record"

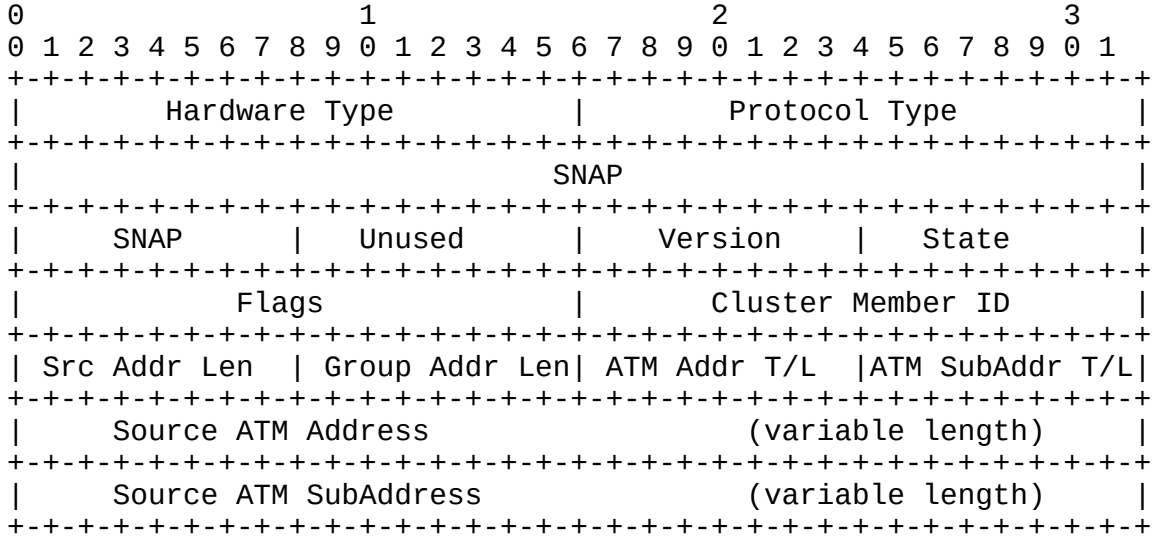
Cache Key Len = 0x04
Orig ID Len = 0x04

See Section B.2.0.2 of [2] for a detailed description of these fields.

5. Detailed State Descriptions

5.1 MARS Server Redirect Entry.

The MARS Server Redirect Entry is used to determine the operational state of a MARS Server in the SG. Each server MUST update it's MARS Server Redirect Entry state at least every 2 minutes, it is RECOMMENDED that it is updated every 1 minute.



Hardware Type
This value is the ATM Forum 'address family number' specified in [3] as 15 decimal (0x000F).

Protocol Type
This field is the protocol type number for the protocol using MARS from [3]. (IPv4 is 0x0800).

Protocol SNAP
This field is the optional protocol SNAP extension to protocol type. This is the mar$pro.snap field from [1].

Version Number
Version Number for this document MUST be set to 0x00.

State
State value is coded as 1 decimal for a MARS Server Redirect Entry.

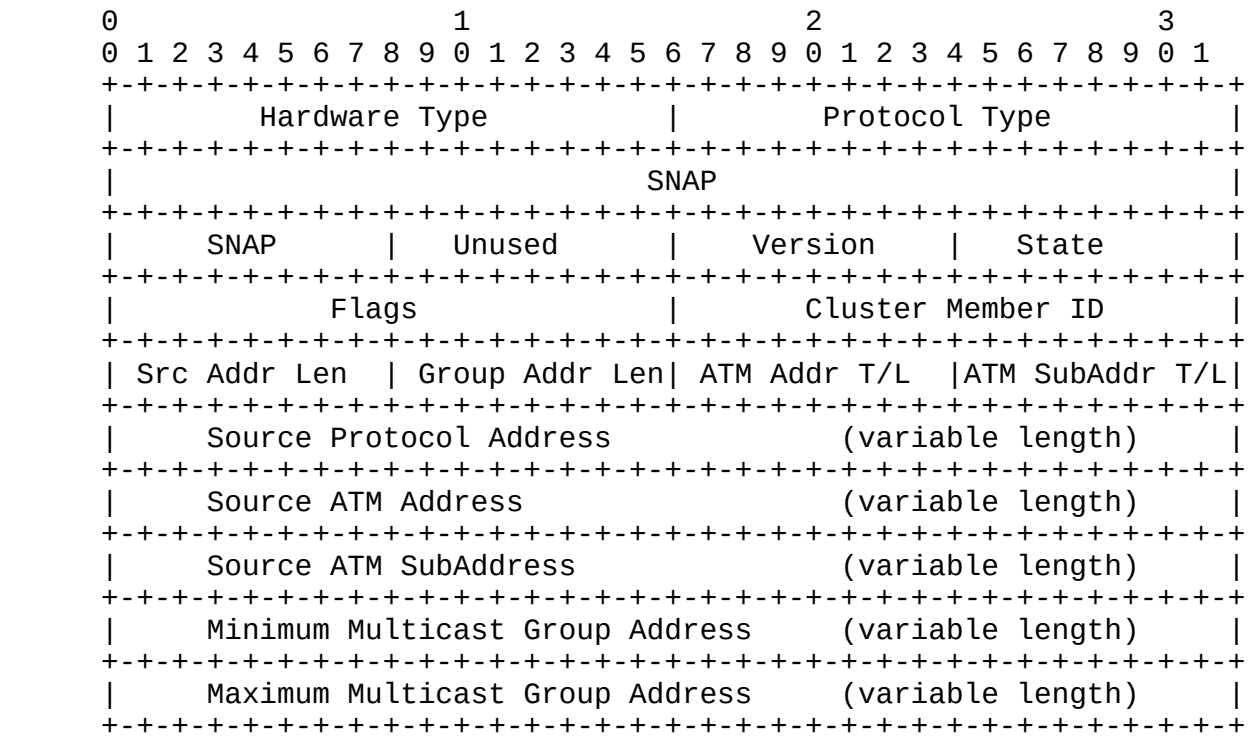
The Flags, Cluster Member ID, Src Addr Len, and Group Addr Len fields are unused and set to zero.

The ATM Addr T/L, ATM SubAddr T/L, Source ATM Address, and Source ATM SubAddress fields define the ATM address for the source of the MARS Server Redirect Entry in the SG. The coding for these fields are the same as described in Section 3 of this document.

Failure to receive two consecutive MARS Server Redirect Entry updates from a given MARS Server in the SG will cause all membership information learned from this server to be flushed. When a valid MARS Server Redirect Entry update is received the source of this update will be placed into the table of backup MARS Servers sent in the MARS_REDIRECT_MAP message. The ordering of servers in the MARS_REDIRECT_MAP will first contain the list of MARS Servers learned via MARS Server Redirect Entry updates in ascending order based on the SCSP Sender ID, followed by any externally configured or learned backup MARS Servers. The format of the MARS_REDIRECT_MAP can be found in Section 5.4.3 of [1].

5.2 MCS Serve/Register request.

The MCS Serve/Register request is used to propagate the registering or servicing of specific groups by an MCS Client within the SG domain. It is similar to an MARS_MSERV request defined in Section 6.2.2 and 6.2.3 of [1]. When a MARS Server in the SG successfully adds a new MCS Client to it's SCVC or adds MCS support for a specific group it MUST send a MCS Serve/Register request to the SG. An MCS Client can only register with one MARS Server in the SG.



Hardware Type
This value is the ATM Forum 'address family number' specified in [3] as 15 decimal (0x000F).

Protocol Type
This field is the protocol type number for the protocol using MARS from [3]. (IPv4 is 0x0800).

Protocol SNAP
This field is the optional protocol SNAP extension to protocol type. This is the mar$pro.snap field from [1].

Version Number
Version Number for this document MUST be set to 0x00.

State
State value is coded as 2 decimal for a MCS Serve/Register request.

Flags
The flags field is used to contain several flags:

```
mar$flags
   Bit 15   - mar$flags.layer3grp
   Bit 13   - mar$flags.register
   Bit 0-7  - mar$flags.sequence
```

The mar$flags.register bit MUST be set the same as in the originating MARS_MSERV request. The mar$flags.layer3grp bit MUST be zero and the mar$flags.sequence bits are of local significance only to the LS.

Cluster Member CMI
This field contains the CMI assigned by the MARS Server which processed the MARS_MSERV request and uniquely identifies the MCS Client in the MARS server cache.

Src Addr Len
This field contains the length of the Source Protocol Address field. For IPv4, the value is 4 if an address is specified. A null (non-existent) address MUST be coded as zero length, and no space allocated for it in the message body.

Group Addr Len
This field contains the length of the Group Protocol Address field. If the register bit in the flags field is set to 1 in the request this field MUST be zero. If the register bit is zero in the flags field the value of this field for IPv4 is 4.

ATM Addr T/L
This field contains the type and length of the Source ATM Address field for the MCS Client that originated the MARS_MSERV request. The type and length encoding is described in Section 3.

ATM SubAddr T/L
This field contains the type and length of the Source ATM SubAddress field for the MCS Client that originated the MARS_MSERV request. The type and length encoding is described in Section 3.

Source Protocol Address
This is the internetwork address for the source of an address binding in a MARS server cache entry. If Src Addr Len is set to zero no storage will be allocated.

Source ATM Address
This is the MCS Client’s ATM address of an address binding in a MARS server cache entry. The address is E.164 or ATM Forum NSAPA.

Source ATM SubAddress
This is the MCS Client's ATM subaddress of an address binding in a MARS server cache entry. The subaddress, if specified, is an ATM Forum NSAPA. If null, no storage will be allocated.

Minimum Multicast Group Address
This is the internetwork address of the lower bound on the range of multicast group addresses for the address binding in a MARS server cache entry. If Group Addr Len is set to zero no storage will be allocated.

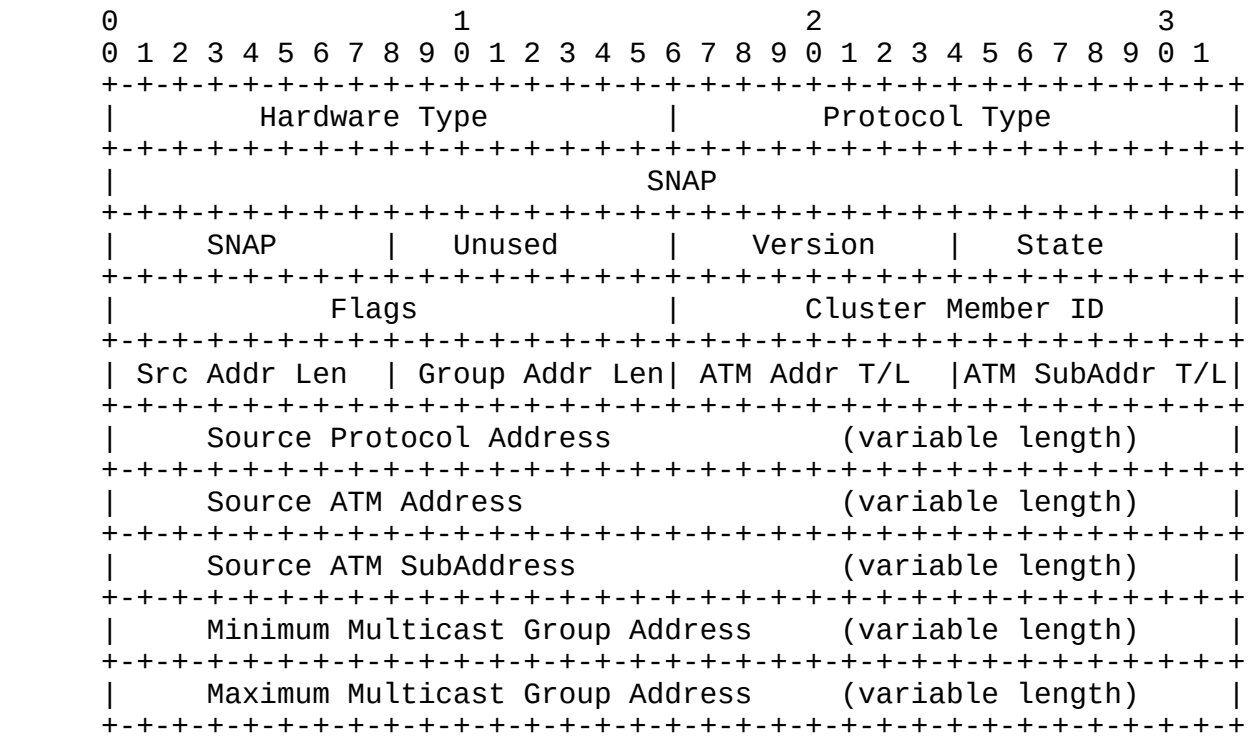
Maximum Multicast Group Address
This is the internetwork address of the upper bound on the range of multicast group addresses for the address binding in a MARS server cache entry. If Group Addr Len is set to zero no storage will be allocated.

An MCS Client can only register with one MARS Server in the SG and is only placed on the SCVC for the MARS Server for which it is registered with.

When a MCS Client Serve/Register request specifying a group address is received by a MARS Server it MUST create a cache entry associated with this client. In addition to adding the cache entry it MUST send out a MARS_MIGRATE message on it's CCVC. This is needed so that clients using a mesh topology can migrate to a server based topology. Details regarding the MARS_MIGRATE message can be found in Section 5.1.6 of [1].

5.3 MARS Client Join/Register request.

The MARS Client Join/Register request is used to propagate the registering or joining of specific group ranges by MARS Clients within the SG domain. It is similar to the MARS_JOIN request defined in Sections 5.2.1 to 5.2.3 of [1]. When a MARS Server in the SG successfully registers a new MARS Client or a registered client joins a specific group address range the MARS Server MUST send a MARS Client Join/Register request to the SG. A MARS Client can only register with one MARS Server in the SG and is placed only on that servers CCVC.



Hardware Type
This value is the ATM Forum 'address family number' specified in [3] as 15 decimal (0x000F).

Protocol Type
This field is the protocol type number for the protocol using MARS from [3]. (IPv4 is 0x0800).

Protocol SNAP
This field is the optional protocol SNAP extension to protocol type. This is the mar$pro.snap field from [1].

Version Number
Version Number for this document MUST be set to 0x00.

State
State value is coded as 3 decimal for a MARS Client Join/Register request.

Flags
The flags field is used to contain several flags:

```
mar$flags
   Bit 15   - mar$flags.layer3grp
   Bit 13   - mar$flags.register
   Bit 0-7  - mar$flags.sequence
```

The mars$flags.layer3grp and mar$flags.register bits MUST be set the same as in the originating MARS_JOIN request. The mar$flags.sequence bits are of local significance only to the LS.

Cluster Member CMI
This field contains the CMI assigned by the MARS Server which processed the MARS_JOIN request and uniquely identifies the MARS Client in the MARS server cache.

Src Addr Len
This field contains the length of the Source Protocol Address field. For IPv4, the value is 4 if an address is specified. A null (non-existent) address MUST be coded as zero length, and no space allocated for it in the message body.

Group Addr Len
This field contains the length of the Group Protocol Address field. If the register bit in the flags field is set to 1 in the request this field MUST be zero. If the register bit is zero in the flags field the value of this field for IPv4 is 4.

ATM Addr T/L
This field contains the type and length of the Source ATM Address field for the MARS Client that originated the MARS_JOIN request. The type and length encoding is described in Section 3.

ATM SubAddr T/L
This field contains the type and length of the Source ATM SubAddress field for the MARS Client that originated the MARS_JOIN request. The type and length encoding is described in Section 3.

Source Protocol Address
This is the internetwork address for the source of an address binding in a MARS server cache entry. If Src Addr Len is set to zero no storage will be allocated.

Source ATM Address
This is the MARS Client’s ATM address of an address binding in a MARS server cache entry. The address is E.164 or ATM Forum NSAPA.

Source ATM SubAddress
This is the MARS Client's ATM subaddress of an address binding in a MARS server cache entry. The subaddress, if specified, is an ATM Forum NSAPA. If null, no storage will be allocated.

Minimum Multicast Group Address
This is the internetwork address of the lower bound on the range of multicast group addresses for the address binding in a MARS server cache entry. If Group Addr Len is set to zero no storage will be allocated.

Maximum Multicast Group Address
This is the internetwork address of the upper bound on the range of multicast group addresses for the address binding in a MARS server cache entry. If Group Addr Len is set to zero no storage will be allocated.

An MARS Client can only register with one MARS Server in the SG and is only placed on the CCVC for the MARS Server for which it is registered with. If the mar$flags.layer3grp is set to 1 than the Minimum and Maximum Multicast Group Addresses MUST be equal for IPv4.

When a MARS Client Join/Register request is sent with the mar$flags.register bit set to 1 all of the servers in the SG will create a cache entry for this client using the information in the request.

When a registered MARS Client issues a MARS_JOIN for a specific group address range a MARS Client Join/Register request MUST be sent to the servers in the SG. The actions taken by each server in the SG depend on previous group membership actions and MCS supported groups.

Each MARS Server MUST perform the necessary redistribution and hole punching algorithms before propagating this request to the CCVC and SCVC on each server. The redistribution and hole punching algorithms used for propagating join requests to the CCVC are the same as defined in Sections 6.1.2 and 6.2.4 of [1]. If the originating MARS_JOIN request is a duplicate of a previously joined range or contains no group address range than a MARS Client Join/Register MUST NOT be sent to the SG.

The redistribution and hole punching algorithms used for propagating join requests as MARS_SJOIN request on a SCVC is the same as Section 6.2.4 except for the following. Only the MARS Servers which contain the registered MCS Clients for the target group ranges should propagate this information to their SCVCs.

5.4 MARS Client Leave/Deregister request.

The MARS Client Leave/Deregister request is used to propagate the deregistering or leaving of specific group ranges by registered MARS Clients within the SG domain. It is similar to the MARS_LEAVE request defined in Sections 5.2.1 to 5.2.3 of [1]. When a MARS Server in the SG successfully deregisters a registered MARS Client or a registered client leaves a specific group address range for which it had joined the MARS Server MUST send a MARS Client Leave/Deregister request to the SG. If a registered MARS Client is unexpectedly removed from the CCVC the MARS Server MUST act as a proxy and send a MARS Client Leave/Deregister request to the SG.

The format and meanings of the fields in a MARS Client Leave/Deregister request are the same as in Section 5.3 except the State is coded as 4 decimal for a MARS Client Leave/Deregister request.

When a MARS Client Leave/Deregister request is sent with the mar$flags.register bit set to 1 all of the servers in the SG receiving this update MUST purge all cache entries for this client.

When a registered MARS Client issues a MARS_LEAVE for a specific group address range a MARS Client LEAVE/Deregister request MUST be sent to the servers in the SG. The actions taken by each server in the SG depend on previous group membership actions and MCS supported groups.

Each MARS Server MUST perform the necessary redistribution and hole punching algorithms before propagating this request to the CCVC and SCVC on each server. The redistribution and hole punching algorithms used for propagating leave requests to the CCVC are the same as defined in Sections 6.1.2 and 6.2.4 of [1]. If the originating MARS_LEAVE request does not correspond to a previously joined range or contains no group address range than a MARS Client Leave/Deregister MUST NOT be sent to the SG.

The redistribution and hole punching algorithms used for propagating leave requests as MARS_SLEAVE requests on a SCVC is the same as Section 6.2.4 except for the following. Only the MARS Servers which contain the registered MCS Clients for the target group ranges should propagate this information to their SCVCs.

5.5 MCS Unserve/Deregister request.

The MCS Unserve/Deregister request is used to propagate the deregistering or unservicing of specific groups by a registered MCS Client within the SG domain. It is similar to an MARS_MUNSERV request

defined in Section 6.2.2 and 6.2.3 of [1]. When a MARS Server in the SG successfully deregisters a registered MCS Client or registered MCS Client stops serving a specific group address range for which it had serviced the MARS Server MUST send a MCS Unserve/Deregister request to the SG. If a registered MCS Client is unexpectedly removed from the SCVC the MARS Server owning the SCVC MUST act as a proxy and send a MCS Unserve/Deregister request to the SG.

The format and meanings of the fields in a MCS Unserve/Deregister request are the same as in Section 5.2 except the State is coded as 5 decimal for a MCS Unserve/Deregister request.

When a MCS Client Unserve/Deregister request is sent with the mar$flags.register bit set to 1 all of the servers in the SG receiving this update MUST purge all cache entries for this client.

When a registered MCS Client issues a MARS_MUNSERV for a specific group address range being served a MCS Client Unserve/Deregister request MUST be sent to the servers in the SG. The members of the SG that receive this update must then clear the cache entry associated with this MCS Client.

In addition to clearing one or more cache entries associated with receiving a MCS Client Unserve/Deregister request each MARS Server in the SG MUST send out a MARS_LEAVE message on it’s CCVC in order for clients to change back to a mesh topology.

6. Security Considerations

There is no mechanism to encrypt the CSA Record MARS Specific Part of the message exchanged between servers. However, there are base SCSP security features in the SCSP Protocol Independent part [2] which can be used to protect against attacks.

Any SCSP MARS is susceptible to Denial of Service (DOS) attacks. A rouge MARS Client can inundate its Server with MARS packets. This is a base MARS problem as currently defined by [1]. A rouge host can also inundate its neighboring SCSP MARS with SCSP packets. However, if the authentication option is used, the SCSP MARS databases will not become corrupted, as the bogus packets will be discarded when the authentication check fails.

Due to the pair wise authentication model of SCSP MARS, the information received from any properly authenticated server is trusted and propagated throughout the server group. Consequently, if security of any SCSP MARS server is compromised, the entire database becomes vulnerable to corruption originating from the compromised server.

References

[1] Armitage, G., "Support for Multicast over UNI 3.0/3.1 based ATM Networks", RFC 2022, November 1996.

[2] Luciani, J., Armitage, G., Halpern, J. and N. Doraswamy, "Server Cache Synchronization Protocol", RFC 2334, April 1998.

[3] Reynolds, J. and J. Postel, "Assigned Numbers", STD 2, RFC 1700, October 1994. See also: http://www.iana.org/numbers.html

[4] Laubach, M., "Classic IP and ARP over ATM", RFC 1577, January 1994.

[5] Bradner, S., "Key Words for use in RFCs to Indicate Requirement Levels," BCP 14, RFC 2119, March 1997.

Acknowledgments

The authors would like to thank Grenville Armitage for his previous distributed MARS work and also the members of the ION working group of the IETF, whose review and discussion of this document has been invaluable.

Authors' Addresses

James V. Luciani
Bay Networks, Inc.
3 Federal Street, BL3-04
Billerica, MA 01821

Phone: +1-508-916-4734
EMail: luciani@baynetworks.com

Anthony M. Gallo
IBM, Networking Hardware Division
Dept. M6LA/B664
P.O. Box 12195
Research Triangle Park, NC 27709

Phone: +1-919-254-9889
EMail: gallo@raleigh.ibm.com

Full Copyright Statement