The ATM Forum

Technical Committee

Multi-Protocol Over ATM Version 1.0

AF-MPOA-0087.000

July, 1997

The information in this publication is believed to be accurate as of its publication date. Such information is subject to change without notice and The ATM Forum is not responsible for any errors. The ATM Forum does not assume any responsibility to update or correct any information in this publication. Notwithstanding anything to the contrary, neither The ATM Forum nor the publisher make any representation or warranty, expressed or implied, concerning the completeness, accuracy, or applicability of any information contained in this publication. No liability of any kind shall be assumed by The ATM Forum or the publisher as a result of reliance upon any information contained in this publication.

The receipt or any use of this document or its contents does not in any way create by implication or otherwise:

- Any express or implied license or right to or under any ATM Forum member company's patent, copyright, trademark or trade secret rights which are or may be associated with the ideas, techniques, concepts or expressions contained herein; nor

- Any warranty or representation that any ATM Forum member companies will announce any product(s) and/or service(s) related thereto, or if such announcements are made, that such announced product(s) and/or service(s) embody any or all of the ideas, technologies, or concepts contained herein; nor

- Any form of relationship between any ATM Forum member companies and the recipient or user of this document.

Implementation or use of specific ATM standards or recommendations and ATM Forum specifications will be voluntary, and no company shall agree or be obliged to implement them by virtue of participation in The ATM Forum.

The ATM Forum is a non-profit international organization accelerating industry cooperation on ATM technology. The ATM Forum does not, expressly or otherwise, endorse or promote any specific products or services.

NOTE: The user's attention is called to the possibility that implementation of the ATM interoperability specification contained herein may require use of an invention covered by patent rights held by ATM Forum Member companies or others. By publication of this ATM interoperability specification, no position is taken by The ATM Forum with respect to validity of any patent claims or of any patent rights related thereto or the ability to obtain the license to use such rights. ATM Forum Member companies agree to grant licenses under the relevant patents they own on reasonable and nondiscriminatory terms and conditions to applicants desiring to obtain such a license. For additional information contact:

The ATM Forum
Worldwide Headquarters
2570 West El Camino Real, Suite 304
Mountain View, CA 94040-1313
Tel: +1-415-949-6700
Fax: +1-415-949-6705

Acknowledgment

Much work went into the development of this specification. It could not have been completed without the ATM Forum Contributions and participation in the working group by many people. In particular, the Editor would like to recognize the following members who made significant contributions to this effort:

Cedell Alexander
Loa Andersson (MPOA Vice Chairman)
Grenville Armitage
Caralyn Brown (Former MPOA Vice Chairman and Editor)
Ross Callon
Andrew Carter
Michael Craren
John Drake
Rob Enns
Norm Finn
Barbara Fox
Eric Gray
Bryan Gleeson
Joel M Halpern
Dave Husak
John Keene
Ali Kujoory
James Luciani
Keith McCloghrie
Drew Perkins
Andrew Smith
Matt Squire
Vijay Srinivasan
Hiroshi Suzuki
George Swallow (MPOA Chairman)
James Watt

The assistance by these members and all who participated in the MPOA Sub-working Group is greatly appreciated.

Andre N. Fredette, Editor

Contents

1. Introduction
[Informative]

Internetwork layer protocols such as IP, IPX and AppleTalk use routers to allow communication across subnet boundaries. Subnets are often built using LAN technologies, Ethernet and Token Ring being the most popular.

The ATM Forum's LAN Emulation LANE provides Emulated LANs (ELANs) that emulate the services of Ethernet and Token Ring LANs across an ATM network. LANE provides many benefits including interoperation with Ethernet and Token Ring hardware and software, allowing a subnet to be bridged across an ATM/LAN boundary. LANE allows a single ATM network to support multiple ELANs. By using ELANs, internetwork layer protocols may operate over an ATM network in essentially the same way that they operate over Ethernet and Token Ring LANs. While LANE provides an effective means for bridging intra-subnet data across an ATM network, inter-subnet traffic still needs to be forwarded through routers.

The IETF Internetworking Over NBMA Networks (ION) Working Group's Next Hop Resolution Protocol (NHRP) [NHRP] and Multicast Address Resolution Server (MARS) [MARS] protocols also allow internetwork layer protocols to operate over an ATM network. These protocols allow the ATM network to be divided into *ION Subnets*, also known as Logical IP Subnets (LISs) or Local Address Groups (LAGs). Routers are required to interconnect these subnets, but NHRP allows intermediate routers to be bypassed on the data path. NHRP provides an extended address resolution protocol that permits Next Hop Clients (NHCs) to send queries between different subnets. Queries are propagated by Next Hop Servers (NHSs) along the routed path as determined by standard routing protocols. This enables the establishment of ATM VCCs across subnet boundaries, allowing inter-subnet communication without requiring routers in the data path.

Even with both LANE and NHRP, a common situation exists where communicating LAN devices are behind LANE edge devices. MPOA allows these edge devices to perform internetwork layer forwarding and establish direct communications without requiring that the LANE edge devices be full function routers.

1.1 What is MPOA?

The goal of MPOA is the efficient transfer of inter-subnet unicast data in a LANE environment. MPOA integrates LANE and NHRP to preserve the benefits of LAN Emulation, while allowing inter-subnet, internetwork layer protocol communication over ATM VCCs without requiring routers in the data path. MPOA provides a framework for effectively synthesizing bridging and routing with ATM in an environment of diverse protocols, network technologies, and IEEE 802.1 Virtual LANs. This framework is intended to provide a unified paradigm for overlaying internetwork layer protocols on ATM. MPOA is capable of using both routing and bridging information to locate the optimal exit from the ATM cloud.

MPOA allows the physical separation of internetwork layer route calculation and forwarding, a technique known as virtual routing. This separation provides a number of key benefits:

1. It allows efficient inter-subnet communication;
2. It increases manageability by decreasing the number of devices that must be configured to perform internetwork layer route calculation;
3. It increases scalability by reducing the number of devices participating in internetwork layer route calculation;
4. It reduces the complexity of edge devices by eliminating the need to perform internetwork layer route calculation.

MPOA provides MPOA Clients (MPCs) and MPOA Servers (MPSs) and defines the protocols that are required for MPCs and MPSs to communicate. MPCs issue queries for shortcut ATM addresses and receive replies from the MPS using these protocols.

MPOA also ensures interoperability with the existing infrastructure of routers. MPOA Servers make use of routers that run standard internetwork layer routing protocols, such as OSPF, providing a smooth integration with existing networks.

1.2 Services Required by MPOA

1. ATM Signaling [UNI 3.0, UNI 3.1, or UNI 4.0].
2. LANE 2.0 [LANE] (as defined in Section 3, Appendix D [LANE]).
3. Next Hop Resolution Protocol [NHRP].

2. Terms and Definitions [Informative]

2.1 Definitions

Control ATM Address:	The address used to set up an SVC to send control packets to an MPOA component. Each MPOA Component has a single Control ATM Address. The Control ATM address may be different from the Data ATM Address.
Control Flow:	A bi-directional flow of Control Messages between two MPOA Components.
Control Messages:	NHRP and MPOA messages, and any other non-data message used by an MPOA Component.
Data ATM Address:	An ATM address used to set up a shortcut. This address may be different from the Control ATM Address.
Data Flow:	A uni-directional flow of data packets to a single destination Internetwork Layer Address.
Data Plane Purge	An NHRP Purge Message sent on the data plane (i.e. a shortcut) by an MPC.
Default Path:	The hop-by-hop path between Routers that a packet would take in the absence of shortcuts, as determined by routing protocols.
DLL Header:	All headers before the Internetwork Layer packet. For example, an Ethernet frame DLL Header includes the Destination MAC Address, the Source MAC Address, and the Ethertype or length and 802.2 LLC/SNAP information.
Edge Device:	A physical device capable of bridging packets between one or more LAN interfaces and one or more LAN Emulation Clients. An Edge Device also contains one or more MPOA Clients allowing it to forward packets across subnet boundaries using an Internetwork Layer protocol.
Egress:	The point where an Outbound Data Flow exits the MPOA System.
Egress Cache:	The collection of Egress Cache Entries in an MPC.
Egress Cache Entry:	Information describing how Internetwork Layer packets from a particular Shortcut are to be encapsulated and forwarded.
Egress MPC:	An MPC in its role at an Egress.
Egress MPS:	The MPS serving an Egress MPC for a particular Outbound Data Flow.
Emulated LAN:	See [LANE].
Flow:	A stream of packets between two entities. Multiple flows may be multiplexed over a single VCC.
Higher Layers:	The software stack above MPOA and LANE, e.g. LLC, bridging, etc.
Inbound Data Flow:	A Data Flow entering the MPOA System.
Inbound Flow:	Data entering the MPOA System.
Ingress:	The point where an Inbound Data Flow enters the MPOA System.
Ingress Cache:	The collection of Ingress Cache Entries in an MPC.
Ingress Cache Entry:	The collection of information dealing with inbound data flows. This information is used to detect flows that may benefit from a shortcut, and, once detected, indicates the shortcut VCC to be used and encapsulation information to be used on the frame.

Ingress MPC:	An MPC in its role at an Ingress.
Ingress MPS:	The MPS serving an Ingress MPC for a particular Inbound Data Flow.
Internetwork Layer:	The protocols and mechanisms used to communicate across subnet boundaries. E.g., IP, IPv6, IPX, DECnet routing, CLNP, AppleTalk DDP, Vines, SNA, etc.
LANE Service Interface:	The interface over which a LEC communicates with an MPC.
MPC Service Interface:	The interface over which an MPC communicates with the Higher Layers.
MPOA Client (MPC):	A protocol entity that implements the client side of the MPOA protocol.
MPOA Component:	An MPC or MPS.
MPOA Device:	A physical device (e.g., router, bridge, or host) that contains one or more MPOA components.
MPOA Host	A host containing one or more LAN Emulation Clients allowing it to communicate using LAN Emulation. An MPOA Host also contains one or more MPOA Clients allowing it to transmit packets across subnet boundaries using an Internetwork Layer protocol.
MPOA Server (MPS):	A protocol entity that implements the server side of the MPOA protocol. An MPOA Server is co-located with a Router.
MPOA System	The set of inter-communicating MPOA Clients and MPOA Servers.
Outbound Data Flow:	A Data Flow exiting the MPOA System.
Outbound Flow:	Data exiting the MPOA System from a Shortcut.
Protocol Address	An internetwork layer address.
Protocol Data Unit (PDU)	A message sent between peer protocol entities.
Router	A device allowing communication across subnet boundaries using an Internetwork Layer protocol. A Router maintains tables for Internetwork Layer packet forwarding and may participate in one or more Internetwork Layer routing protocols for this purpose. A Router forwards packets between subnets in accordance with these tables. When referred to in this specification, a Router contains , one or more LAN Emulation Clients, one or more MPOA Servers, one or more Next Hop Servers, zero or more Next Hop Clients, and zero or more MPOA Clients.
Routing Protocol:	A protocol run between Routers to exchange information used to allow computation of routes. The result of the routing computation will be one or more next hops.
Service Data Unit (SDU)	A message sent between an entity and its service user or service provider.
Shortcut	An ATM VCC used to forward data packets in lieu of the default routed path.
Target:	An Internetwork Layer Address to which a Shortcut is desired.
Tag:	A 32 bit opaque pattern that an Egress MPC may provide to an Ingress MPC. If a Tag is provided to an Ingress MPC by an Egress MPC, the Ingress MPC must include the tag in the MPOA packet header for packets sent to the given MPC for the given internetwork destination.

2.2 Acronyms and Abbreviations

ARP	Address Resolution Protocol
B-LLI	Broadband Low Layer Information
BCOB-C	Broadband Bearer Connection Oriented Service Type C
BCOB-X	Broadband Bearer Connection Oriented Service Type X
BUS	Broadcast and Unknown Server
CIE	NHRP Client Information Element
CPCS-PDU	Common Part Convergence Sub-layer Protocol Data Unit
DLL	Data Link Layer
ELAN	Emulated LAN
IE	Information Element
IETF	Internet Engineering Task Force
InATMARP	Inverse ATM Address Resolution Protocol
ION	Internetworking Over NBMA (Non-Broadcast Multi-Access)
IP	Internet Protocol
IPX	Internetwork Packet Exchange
L3	Internetwork Layer
LANE	LAN Emulation
LEC	LAN Emulation Client
LECS	LAN Emulation Configuration Server
LES	LAN Emulation Server
LIS	Logical IP Subnet
LLC	Logical Link Control
MARS	Multicast Address Resolution Server
MPC	MPOA Client
MPOA	Multiprotocol Over ATM
MPS	MPOA Server
MTU	Maximum Transmission Unit
NBMA	Non-Broadcast Multi-Access (e.g. ATM, Frame Relay)
NHC	Next Hop Client
NHRP	Next Hop Resolution Protocol
NHS	Next Hop Server
NLSP	NetWare Link State Protocol
OSPF	Open Shortest Path First
PCR	Peak Cell Rate
PDU	Protocol Data Unit
QoS	Quality of Service
RIP	Routing Information Protocol
RSVP	Resource ReSerVation Protocol
SCSP	Server Cache Synchronization Protocol
SDU	Service Data Unit
SNAP	SubNetwork Attachment Point
SVC	Switched Virtual Channel Connection
TLV	Type-Length-Value Encoding
TTL	Time To Live
VCC	Virtual Channel Connection

2.3 Normative Statements

The normative sections of this specification are Section 4, Section 5, and all Annexes. Throughout these normative sections, normative statements are used as follows:

Table 1. Normative Statements

Statement	Verbal Form
Requirement	must/must not
Recommendation	should/should not
Permission	may

The term “may” is used to indicate that a particular procedure is allowed but not required. It is an implementation choice. “may” is also used to indicate allowed behaviors that must be accommodated.

3. MPOA Description [Informative]

MPOA is designed with a client/server architecture. MPOA Clients (MPC) and their MPOA Server(s) (MPS) are connected via LANE.

3.1 MPOA Components

There are two types of MPOA logical components: MPC and MPS.

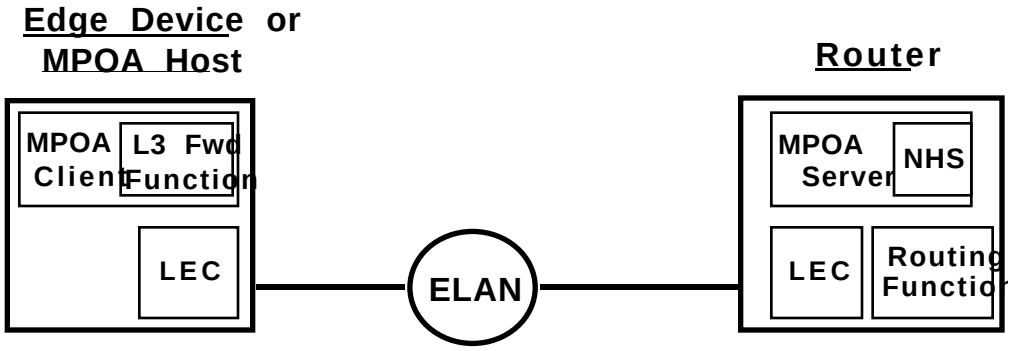


Figure 1 The Components in an MPOA System.

3.1.1 MPOA Client (MPC)

The primary function of the MPC is to source and sink internetwork shortcuts. To provide this function, the MPC performs internetwork layer forwarding, but does not run internetwork layer routing protocols.

In its ingress role, an MPC detects flows of packets that are being forwarded over an ELAN to a router that contains an MPS. When it recognizes a flow that could benefit from a shortcut that bypasses the routed path, it uses an NHRP-based query-response protocol to request the information required to establish a shortcut to the destination. If a shortcut is available, the MPC caches the information in its ingress cache, sets up a shortcut VCC, and forwards frames for the destination over the shortcut.

In its egress role the MPC receives internetwork data frames from other MPCs to be forwarded to its local interfaces/users. For frames received over a shortcut, the MPC adds the appropriate DLL encapsulation and forwards them to the higher layers (e.g., a bridge port or an internal host stack). The DLL encapsulation information is provided to the MPC by an egress MPS and stored in the MPC's egress cache.

An MPC can service one or more LECs and communicates with one or more MPSs.

3.1.2 MPOA Server (MPS)

An MPS is the logical component of a router that provides internetwork layer forwarding information to MPCs. It includes a full NHS as defined in [NHRP] with extensions as defined in this document. The MPS interacts with its local NHS and routing functions to answer MPOA queries from ingress MPCs and provide DLL encapsulation information to egress MPCs.

An MPS converts between MPOA requests and replies, and NHRP requests and replies on behalf of MPCs.

3.1.3 Examples of MPOA Enabled Devices

- MPOA Edge Device (including the MPC, the LEC and a bridge port.)
- MPOA Host (including the MPC, the LEC and an internal host stack.)

- Router (including the MPS, which in turn includes an NHS; the LEC and the routing function)

There are other possibilities to create MPOA enabled devices, e.g. co-locating the MPS and MPC in the router and thereby creating a device that is capable of internetwork routing/forwarding and detecting flows and creating ATM shortcuts for these flows.

3.1.4 Relationship Between LECs, MPOA Components, and MPOA Devices

As shown in Figure 2, there may be one or more MPCs in an edge device and one or more LECs associated with an MPC; however, a given LEC may be associated with one and only one MPC.

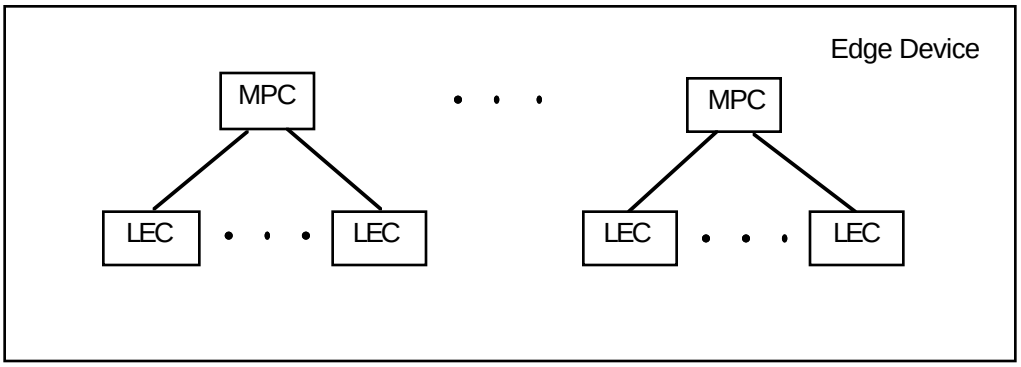


Figure 2 Relationship Between LECs, MPCs, and Edge Devices

Similarly, as shown in Figure 3, there may be one or more MPSs in a router and one or more LECs associated with an MPS; however, a given LEC may be associated with one and only one MPS.

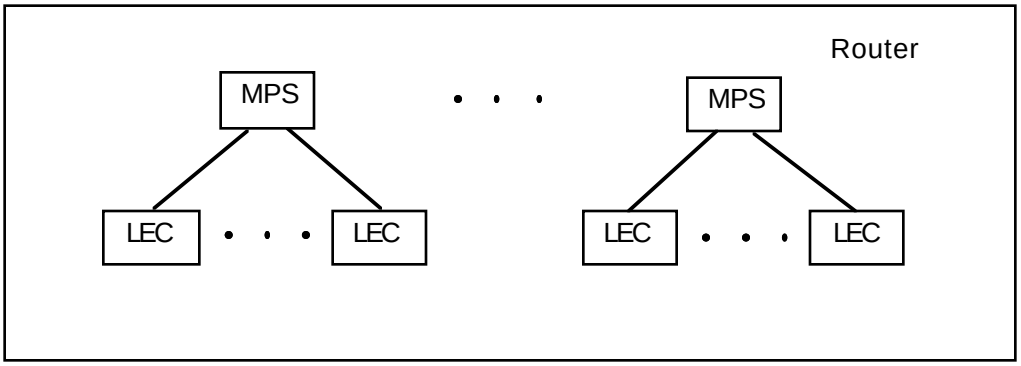


Figure 3 Relationship Between LECs, MPSs, and Routers

3.2 Control and Data Flows

The MPOA solution involves a number of information flows, shown in Figure 4, that can be categorized as MPOA control flows and MPOA data flows.

By default, all control and data flows are carried over ATM VCCs using LLC/SNAP [RFC 1483] encapsulation. Configuration flows use the formats described in [LANE]. More detailed discussion of the information flows is contained in the area descriptions in Section 3.3.

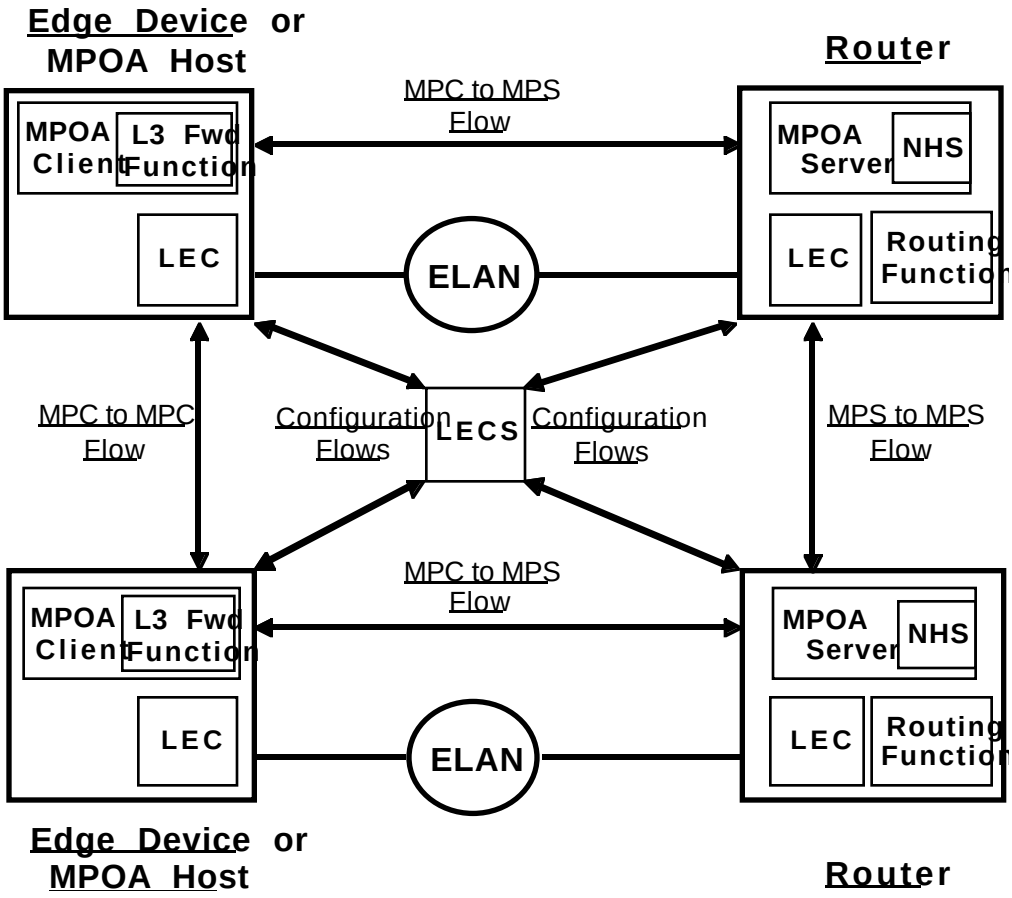


Figure 4 Information Flows in an MPOA System

3.2.1 MPOA Control Flows

3.2.1.1 Configuration Flows

By default, MPSs and MPCs communicate with the LAN Emulation Configuration Server (LECS) to retrieve configuration information.

3.2.1.2 MPC-MPS Control Flows

MPC-MPS control flows are used for MPC cache management. The MPOA Resolution Request/Reply allows the ingress MPC to obtain shortcut information. The ingress MPS may trigger the ingress MPC to make a request by sending the MPOA Trigger Message. The MPOA Cache Imposition Request/Reply allows the egress MPS to give the egress MPC egress cache information. Finally, either the egress MPC or an MPS may send a Purge message if it discovers that cached information has become invalid.

3.2.1.3 MPS-MPS Control Flows

MPS-MPS control flows are handled by standard internetwork layer routing protocols and NHRP. MPOA does not define any new MPS-MPS protocols. MPOA requires no new replication techniques and relies upon the standard replication techniques provided by LANE and internetwork layer routing protocols.

3.2.1.4 MPC-MPC Control Flows

An egress MPC may send a data plane purge to an ingress MPC if it receives misdirected packets from that MPC. This message causes the ingress MPC to invalidate its erroneous cache information.

3.2.2 MPOA Data Flows

3.2.2.1 MPC-MPC Data Flow

MPC-MPC flows are used primarily for the transfer of data between MPCs over MPOA shortcut VCCs.

3.2.2.2 MPC-NHC Data Flows

An MPC may send unicast data to an NHC and an NHC may send unicast data to an MPC.

3.3 MPOA Operations

MPOA performs the following operations:

Configuration	Obtaining the appropriate configuration information.
Discovery	MPCs and MPSs learning of each others' existence.
Target Resolution	Determining the mapping of a Target to an egress ATM address, an optional Tag, and a set of parameters used to set up a Shortcut VCC to forward packets across subnet boundaries.
Connection Management	Creating, maintaining, and terminating VCCs for the purpose of transferring control information and data.
Data Transfer	Forwarding internetwork layer data across a Shortcut.

3.3.1 Configuration

MPCs and MPSs each require configuration. By default, MPOA components retrieve their configuration parameters from the LECS. MPOA components must be capable of configuration via the LECS, although they may be administered to obtain their configuration by some other means. Other methods for obtaining configuration may include manipulation of the MPOA MIB, or through unspecified mechanisms.

3.3.2 Discovery

To reduce operational complexity, MPOA components automatically discover each other using extensions to the LANE LE_ARP protocol that carry the MPOA device type (MPC or MPS) and ATM address. This information is discovered dynamically and used as needed. This information may change and must be periodically verified.

MPCs are not NHCs and do not register host internetwork layer addresses with NHSs using NHRP Registration.

3.3.3 Target Resolution

MPOA target resolution uses an extended NHRP Resolution Request protocol to allow MPCs to determine the ATM address for the end points of a shortcut. In the following subsections, the protocol is described from the perspectives of the ingress MPC, the ingress MPS, the egress MPS, and the egress MPC.

3.3.3.1 Ingress MPC Perspective

An ingress MPC learns the MAC addresses of MPSs attached to its ELANs from the device type TLVs in LE_ARP responses. The MPC is required to perform flow detection, based on internetwork layer destination address, on packets destined for these learned MAC addresses. Additionally, an MPC is permitted to perform other types of flow detection. An example of this is if the MPC is co-located with an MPOA host, it may "detect flows" based on

higher-layer information readily available from the host. In addition, the MPC should issue a request to an MPS from which it has received an MPOA Trigger (described in Section 4.7.2).

Default forwarding for the MPOA System is via routers. When an MPC becomes aware of a particular traffic flow that might benefit from a shortcut, the ingress MPC needs to determine the ATM address associated with the egress device. Note that the terms ingress and egress apply even if both MPCs are part of MPOA hosts. To obtain the ATM address for a shortcut, the ingress MPC sends an MPOA Resolution Request to the appropriate ingress MPS. When this MPS is able to resolve the MPOA Resolution Request, a reply is returned to the ingress MPC that contains an ATM address of the egress device.

The reply may contain information in addition to the requested ATM address. An example of information that may be included is encapsulation/tagging to be used for data sent on this shortcut. Note that NHRP is specified in such a way that only that information requested by the Resolution Request initiator may be included in the reply.

3.3.3.2 Ingress MPS Perspective

The ingress MPS processes MPOA Resolution Requests sent by local MPCs. The ingress MPS can answer the request if the destination is local, otherwise it re-originates the request along the routed path through its local NHS. The ingress MPS uses its internetwork layer address as the source protocol address in the re-originated request. . This ensure that the reply is returned to the originating MPS. The MPS copies all other fields from the MPOA Resolution request. In particular, the MPC's data ATM address is used as the source NBMA address and all TLVs are copied. The MPS generates a new Request ID for the re-originated request. The MPS must set the S bit in the re-originated request to zero so that downstream NHSs do not cache the association of the resulting internetwork layer and ATM addresses.

On receiving a reply to this re-originated request, the ingress MPS restores the Request ID field and source protocol address to the original values and returns an MPOA Resolution Reply to the ingress MPC.

3.3.3.3 Egress MPS Perspective

When an NHRP Resolution Request targeted for a local MPC arrives at the egress MPS serving that MPC (the MPS, in this case, is the NHRP "authoritative responder"), the egress MPS sources an MPOA Cache Imposition Request.

The MPOA Cache Imposition Request is generated by the egress MPS and sent to the egress MPC. It is part of a cache management protocol that serves multiple purposes; the MPOA Cache Imposition Request provides encapsulation and state maintenance information needed by the egress MPC, while the MPOA Cache Imposition Reply provides status, address and ingress tagging information needed by the egress MPS to formulate the NHRP Resolution Reply.

After receiving the MPOA Cache Imposition Reply from the egress MPC, the egress MPS sends an NHRP Resolution Reply toward the request originator. Additional information requested by the ingress MPC (and included in the MPOA Cache Imposition Request and MPOA Cache Imposition Reply messages) must be included in the NHRP Resolution Reply as well.

3.3.3.4 Egress MPC Perspective

The egress MPC must send an MPOA Cache Imposition Reply for every MPOA Cache Imposition Request. To formulate its reply, the MPC must determine if it has the resources necessary to maintain the cache entry and potentially receive a new VCC. If the MPOA Cache Imposition Request is an update of an existing egress cache entry, the resources are likely available. If the MPC cannot accept either the cache entry or the VCC that will likely result from a positive reply, it sets the appropriate error status and returns the MPOA Cache Imposition Reply to the MPS. If it can accept this cache entry, the MPC inserts an ATM address and, if present, may modify the MPOA Egress Cache Tag Extension to be used by the ingress MPC in connection with this shortcut, sets a success status, and sends the MPOA Cache Imposition Reply to the egress MPS.

In some configurations, it is possible for an egress MPC to receive conflicting next hop forwarding instructions for the same source ATM address and destination internetwork layer address pair (as described in Appendix IV). If this conflict occurs, the MPC must take one of the following actions so that packets are forwarded properly:

1. If there is an MPOA Egress Cache Tag Extension present, the egress MPC may include an appropriate tag (unique to the source-destination ATM address pair and internetwork layer destination address) in the MPOA Cache Imposition Reply.
2. Return a distinct destination ATM address in the MPOA Cache Imposition Reply (thus forcing the requesting MPC to open a new VCC).
3. Refuse the cache imposition - indicating in an MPOA Cache Imposition Reply either that additional shortcuts are not possible or that a shortcut for this particular flow is refused.

While the primary technical reason for including a tag is to solve the egress cache conflict referenced above, it is important to note that tags can also be used to optimize the egress cache lookup. This optimization can be achieved by providing an index into the egress cache as the tag. When the tag is an index into the cache, the cache search is reduced to a direct cache lookup.

3.3.4 Connection Management

MPOA components establish VCCs between each other as necessary to transfer data and control messages over an ATM network. For the purpose of establishing control VCCs, MPOA components learn of each others existence by the discovery process described in Section 3.3.2. For the purpose of establishing data VCCs, MPOA components learn of each others existence by the resolution process described in Section 3.3.3.

3.3.5 Data Transfer

The primary goal of MPOA is the efficient transfer of unicast data. Unicast data flow through the MPOA System has two primary modes of operation: the default flow and the shortcut flow. The default flow follows the routed path over the ATM network. In the default case, the MPOA edge device acts as a layer 2 bridge. Shortcuts are established by using the MPOA target resolution and cache management mechanisms.

When an MPC has an Internetwork protocol packet to send for which it has a shortcut, the MPOA edge device acts as an internetwork level forwarder and sends the packet over the shortcut.

3.4 Routing Protocol Interaction

Routing information is supplied to MPOA via the NHS and its associated routing function. MPSs interact with NHSs to initiate and answer resolution requests. Ingress and egress NHSs (associated with MPSs) must maintain state on NHRP Resolution Requests that they have initiated or answered so that they can update forwarding appropriately if routing information changes. MPSs receive these updates from their co-located router/NHS and update or purge relevant MPC caches as appropriate. Interactions between an NHS and internetwork layer routing protocols are beyond the scope of this document.

3.5 NHRP/ION Interaction

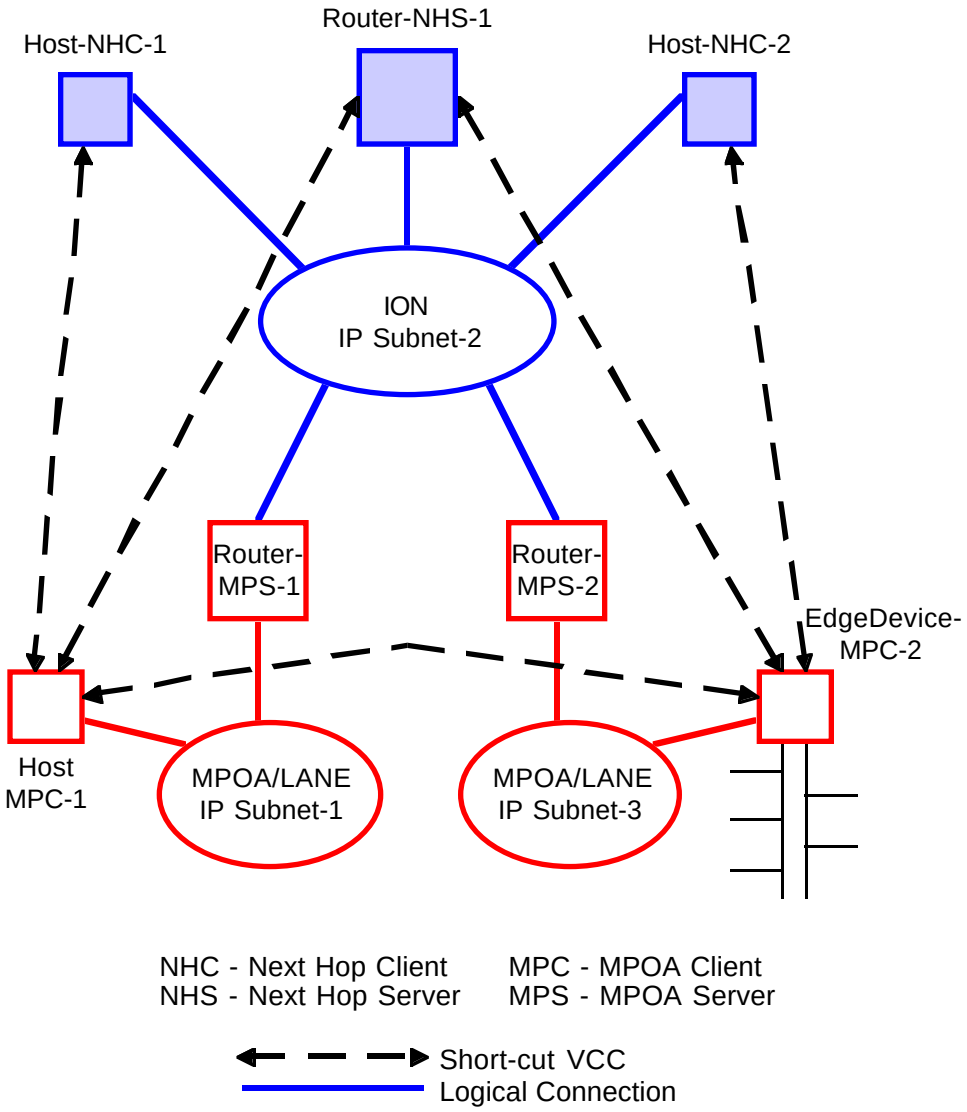


Figure 5 MPOA / ION Interaction

MPOA supports interoperation between MPOA devices and NHRP-only devices. To the NHRP devices in the ION domain, MPCs appear to be standard NHCs. As shown in Figure 5, unicast shortcuts can be established between MPOA devices and NHRP-only devices in different IP subnets. In particular, unicast shortcuts can be established between MPOA hosts and NHRP-capable hosts and routers, and unicast shortcuts can be established between MPOA edge devices and NHRP-capable hosts and routers. The one restriction related to MPOA/NHRP interoperability is that MPOA devices and NHRP (and Classical IP [CLIP]) devices must be on different subnets because intra-subnet unicast and multicast between MPOA and ION devices are not specified in this document.

3.6 A Day in the Life of a Data Packet

A packet enters the MPOA System at the ingress MPC (MPC 1). The decision process that takes place relative to each inbound packet at an MPC is outlined in Figure 8. By default, the packet is bridged via LANE to a router. If the packet follows the default path, it leaves the MPOA System via the ingress MPC's internal LEC Service Interface. However, if this packet is part of a flow for which a shortcut has been established, the ingress MPC strips

the DLL encapsulation from the packet and sends it via the shortcut. The MPC may be required to prefix the packet with tagging information (provided to the MPC via target resolution process - Section 3.3.3) prior to sending it via the shortcut.

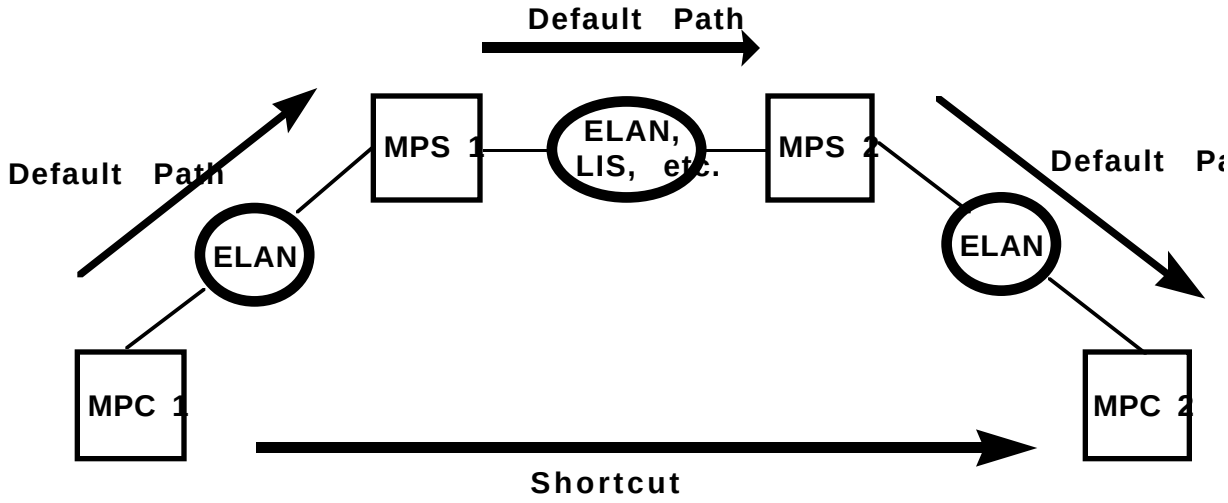


Figure 6 Example of a Day in the Life of a Packet

If no flow has been detected previously, each packet being sent to an MPS is tallied by internetwork layer destination address as it is being sent via LANE. When a threshold (given as a number of packets for a single internetwork layer address in a fixed period of time) is exceeded, the MPC is required to send an MPOA resolution request to obtain the ATM address to be used for establishing a shortcut to a specific downstream element - most likely an egress MPC (e.g. MPC 2).

On arriving via shortcut at the egress MPC, a packet is examined and either a matching egress cache entry is found or the packet is dropped. If a match is found, the packet is encapsulated using the information in the egress cache, and it is forwarded to the higher layer.

Appendix I provides example scenarios for the data and control flows.

4. MPOA Specification
[Normative]

MPOA uses a protocol based on the Next Hop Resolution Protocol [NHRP] to manage caches and establish shortcuts. This section describes the MPOA protocol including all configuration parameters, initial and operating states, and packet processing procedures. Section 4.1 describes all MPS and MPC configuration parameters and procedures. Section 4.2 describes the procedure by which MPOA components automatically discover each other. Section 4.3 describes the generic retry mechanism that MPOA components must use when retrying requests. Section 4.4 describes detailed MPC Behavior, and section 4.5 describes detailed MPS Behavior. Section 4.6 describes a Keep-Alive protocol by which MPCs detect the death of MPSs to ensure cache consistency. Section 4.7 describes cache maintenance. Finally, Section 4.8 describes connection management.

4.1 Configuration Parameters

Sections 4.1.1 - 4.1.2.2 describe the MPOA configuration parameters and constants. The granularity for the parameters and constants described is the same as is given in the tables.

4.1.1 MPS Configuration

Most MPS configuration information (such as what ELANs to operate over) can be derived from the underlying router configuration. Some additional configuration information is specific to the MPS.

4.1.1.1 MPS Parameters

The following parameters apply to each MPS:

Variable	Name	Description and Values
MPS-p1	Keep-Alive Time	The MPS must transmit MPOA Keep-Alives every MPS-p1 seconds. Minimum=1 second, Default=10 seconds, Maximum=300 seconds.
MPS-p2	Keep-Alive Lifetime	The length of time an MPC may consider a Keep-Alive valid in seconds. Minimum=3 seconds, Default=35 seconds , Maximum=1000 seconds (MPS-p2 must be at least three times MPS-p1)
MPS-p3	Internetwork-layer Protocols	The set of protocols for which MPOA resolution is supported. Default ={}.
MPS-p4	MPS Initial Retry Time	Initial retry time used by the MPOA retry mechanism. Minimum=1 second, Default=5 seconds, Maximum=300 seconds
MPS-p5	MPS Retry Time Maximum	Maximum retry time used by the MPOA retry mechanism. Minimum=10 seconds, Default=40 seconds, Maximum=300 seconds
MPS-p6	MPS Give Up Time	Minimum time to wait before giving up on a pending resolution request. Minimum = 5 seconds, Default = 40 seconds, Maximum = 300 seconds
MPS-p7	Default Holding Time	The default Holding Time used in NHRP Resolution Replies. An egress MPS may use local information to determine a more appropriate Holding Time. Minimum=1 Minute, Default=20 Minutes, Maximum=120 Minutes

4.1.1.2 MPS Constants

The following constants are used by MPSs.

Constant	Name	Description and Values
MPS-c1	Retry Time Multiplier	Value: 2

4.1.2 MPC Configuration

4.1.2.1 MPC Parameters

The following parameters apply to each MPC:

Variable	Name	Description and Values
MPC-p1	Shortcut-Setup Frame Count	See parameter MPC-p2. Minimum=1, Default=10, Maximum=65535.
MPC-p2	Shortcut-Setup Frame Time	If an MPC forwards at least MPC-p1 frames to the same target within any period MPC-p2 via the default forwarding path, it should initiate the procedure to establish a shortcut.. The MPC-p1 and MPC-p2 parameters specify a default mechanism for automatically detecting flows in the absence of other information. Other mechanisms (e.g. RSVP) may be used in specific cases to override this default. Minimum=1 second, Default=1 second, Maximum=60 seconds.
MPC-p3	Flow-detection Protocols	A set of protocols on which to perform flow detection. Default ={}.
MPC-p4	MPC Initial Retry Time	Initial retry time used by the MPOA retry mechanism. Minimum=1 second, Default=5 seconds, Maximum=300 seconds
MPC-p5	MPC Retry Time Maximum	Maximum retry time used by the MPOA retry mechanism. Minimum=10 seconds, Default=40 seconds, Maximum=300 seconds
MPC-p6	Hold Down Time	Minimum time to wait before reinitiating a failed resolution attempt. This is usually set to a value greater than MPC-p5. Minimum=30 seconds, Default=MPC-p5*4, Maximum=1200 seconds

4.1.2.2 MPC Constants

The following constants are used by MPCs.

Constant	Name	Description and Values
MPC-c1	Retry Time Multiplier	Value: 2
MPC-c2	Initial Keep-Alive Lifetime	Keep-Alive Lifetime to use before the first Keep-Alive message is received. Value: 60 seconds.

4.2 Device Discovery

Discovery of control addresses of MPOA components is an essential part of the MPOA system. It is necessary for the MPOA Client to know the MAC and ATM addresses of local MPOA Servers so that MPOA Requests may be sent. The MPOA Server must know if an NHRP Request resolves to the ATM address of an MPOA Client so that a cache imposition may be sent. Finally, MPSs sharing an ELAN must be able to discover each other to facilitate the forwarding of NHRP messages. To this end, an MPOA Device Type TLV, defined in Section 5.2.3, is included in the following LANE messages:

- LE_REGISTER_REQUEST
- LE_REGISTER_RESPONSE
- LE_ARP_REQUEST
- LE_ARP_RESPONSE
- Targetless LE_ARP_REQUEST

The information carried in the MPOA Device Type TLV includes the type of device (MPS, MPC, MPS/MPC, or non-MPOA), MPS MAC addresses (if the type is MPS), and the respective control addresses. Each MPOA component must register the MPOA Device Type TLV with each LEC on which it is configured.

4.2.1 Register Protocol

If a LEC is being served by an MPOA Server or Client, it must include the MPOA Device Type TLV in the Register Request for the relevant MAC addresses. The LEC indicates the type of device in the TLV.

If the LES has no existing entry for the MAC-ATM binding, it must register the new MAC-ATM binding with MPOA Identification information in its address table. However, if the LES has an existing entry for the specified MAC-ATM binding, it must overwrite any existing MPOA Identification with the new MPOA Identification.

If the status of an MPOA device changes, it should request each of its served LECs to send an Unregister Request to the LES for each registered MAC address. After unregistering, the LEC should send a Register Request with the new set of TLVs to the LES for each MAC address.

4.2.2 Address Resolution Protocol

Address resolution requests and replies sent by a LEC with an ATM address that is associated with an MPOA device must include the MPOA Device Type TLV (as described in Section 5.2.3). The TLV will indicate whether the sending MPOA device is an MPOA Server, an MPOA client, or both. A LEC receiving an address resolution request or response must update its MAC-ATM binding entry to reflect the MPOA Identification TLV.

If the status of an MPOA device changes, it should request each of its served LECs to send a TARGETLESS_LE_ARP_REQUEST to the LES for MAC addresses previously included in an address resolution response with the new set of TLVs.

4.2.3 Implications for Co-Located MPS, MPC and Non-MPOA Devices

A device may have one or more MAC addresses on a LANE LEC which are associated with an MPS or MPC in the device, and one or more MAC addresses which are not. For example, a device may be both a router and a bridge. It might have a router MAC address associated with an MPS, but still respond to other MAC addresses which it is bridging. Those bridged MAC addresses would not be associated with the MPS. As another example, a bridge with an MPC might or might not want to associate the MPC with the MAC address it uses for SNMP traffic to the bridge, itself.

Any MPOA device which has a LEC that has a set of MAC addresses associated with an MPC, and a set of MAC addresses not associated with any MPOA capability must have separate LEC ATM addresses (or sets of ATM addresses) associated with those two sets of MAC addresses. No LEC ATM address given out with a non-MPOA MAC address can also be given out for an MPC MAC address. This means that traffic for MPC MAC addresses and non-MPOA MAC addresses cannot share the same VCC, but must be carried on separate VCCs. This is to allow

other MPOA devices which are performing learning of MAC to VCC/ATM address bindings from LANE data frames, to determine the correct MPOA capabilities of the MAC addresses learned in this manner.

A router with a LEC that has a set of MAC addresses associated with an MPS, is not required to use a separate LEC ATM address for the MPS MAC addresses. Instead it may share the LEC ATM address with MPC or non-MPOA MAC addresses. It may also choose to use a separate LEC ATM address and not share.

The ability to share is achieved by including the list of MPS MAC addresses in a Device Type TLV in LANE control messages sent pertaining to the LEC ATM address serving the set of MPS MAC addresses. For Requests this means that the SOURCE-ATM -ADDRESS field contains the LEC ATM address, for Responses this means that the TARGET-ATM-ADDRESS field contains the LEC ATM address.

Note that if sharing is used, the list of MPS MAC addresses will be included in LANE control messages (as defined in Section 4.2) issued for or on behalf of an MPC or non-MPOA MAC address served by that shared LEC ATM address.

By including an explicit list of MPS MAC addresses in the Device Type TLV, other MPOA devices which are performing learning of MAC to VCC/ATM address bindings from LANE data frames can determine the correct MPOA capabilities of the MAC addresses learned in this manner, even if a separate LEC ATM address is not used for MPS MAC addresses. All MAC addresses learned that are not in the MPS list, are associated with either an MPC or as not being MPOA capable, as determined by the capability associated with the VCC.

A device that is performing learning of MAC to VCC/ATM address bindings from LANE data frames is required to perform at least one LE-ARP on at least one MAC address for every ATM address to which it has a Data Direct VCC to discover the MPOA capabilities of the device at the other end of the VCC. It does this to associate the correct MPOA capabilities with the MAC addresses it learns in this way. This rule means that, if a LEC accepts a Data Direct VCC from another device, and it has no LE_ARP cache entry for that ATM address, and it then receives data frames from that device, it must not simply learn the source MAC addresses from the frames and populate its LE_ARP cache with that learned information. It may populate its LE_ARP cache in this manner, but it must, in addition, issue an LE_ARP for one of those learned MAC addresses, and receive the reply, to find out whether the associated ATM address is or is not associated with an MPC or MPS.

Note that these arguments do not prevent an MPS and MPC in the same device from sharing a common ATM address for their non-LANE control connections. The MPOA packet formats are chosen so that messages for those two entities cannot be confused. Furthermore, an MPS or MPC may share an ATM address with one used by one or more of its LECs for MPS/MPC associated MAC addresses.

4.2.4 Change of Device Status

Note that, in order for the MPOA-aware devices to be able to trust just one LE_ARP request or response, a device must be careful when changing its configuration to start or stop an MPS or MPC, start or stop an associated LEC, or associate a LEC with, or disassociate a LEC from, an MPS or MPC. There are a number of ways to notify other MPOA-aware devices of a change in the configuration of an MPS, MPC, or associated LEC:

1. A LEC or MPC or MPS may terminate ELAN membership and/or cease operations for a significantly longer period than the maximum LE_ARP time-out period (5 minutes in [LANE]). This ensures that information about the device will age out.
2. A LEC may send one or more Targetless LE_ARP_REQUESTs to update other LECs' MPS/MPC associations. Such a Targetless LE_ARP_REQUEST may be sent more than once for reliability, and may be sent each time an indication (such as an MPOA request sent to a device in which the MPC or MPS is no longer active) is received that some other device is confused.
3. An MPC (MPS) may respond to an MPOA request with error code 0x86 (0x87), meaning, "this device is no longer an MPC (MPS)."

Of all these methods, only the first is a reliable means for a device to notify all other devices of a change in its configuration.

4.3 MPOA Retry Mechanism

MPOA Requests may be retried if a response is not received within a reasonable amount of time. The following retry mechanism is defined for MPOA components when retrying request messages. While retrying a failed request, the MPOA component must use the same Request Id that was used in the initial request (if appropriate) . Retries for failed requests must use the retry procedure as follows: The retry procedure includes an initial time-out of MPS-p4/MPC-p4 seconds, a retry multiplier of MPS-c1/MPC-c1, and a cumulative maximum time-out of MPS-p5/MPC-p5. When an MPOA component sends a request, it sets a RetryTimer to the value of MPS-p4/MPC-p4 seconds. If a corresponding reply is not received before RetryTimer seconds have elapsed, the MPOA component may send a new request (retry). Each time a retry is sent, the RetryTimer is set to RetryTimer*MPS-c1/MPC-c1. If the value of RetryTimer exceeds the Retry Time Maximum (MPS-p5/MPC-p5), the request is considered to have failed.

4.4 Detailed MPC Behavior

The MPC lies between a LANE LEC and its higher layers. Each LEC for which MPOA is enabled is associated with exactly one MPC. Each MPC serves a set of one or more LECs, and has a single MPC control ATM address. This address may be the same as an ATM address used by one or more of its LECs. If there are multiple MPCs within a device, each must serve a disjoint set of LECs, and must use different MPC control ATM addresses. For example, Figure 7 shows an MPOA edge device with a single MPC and two LECs.

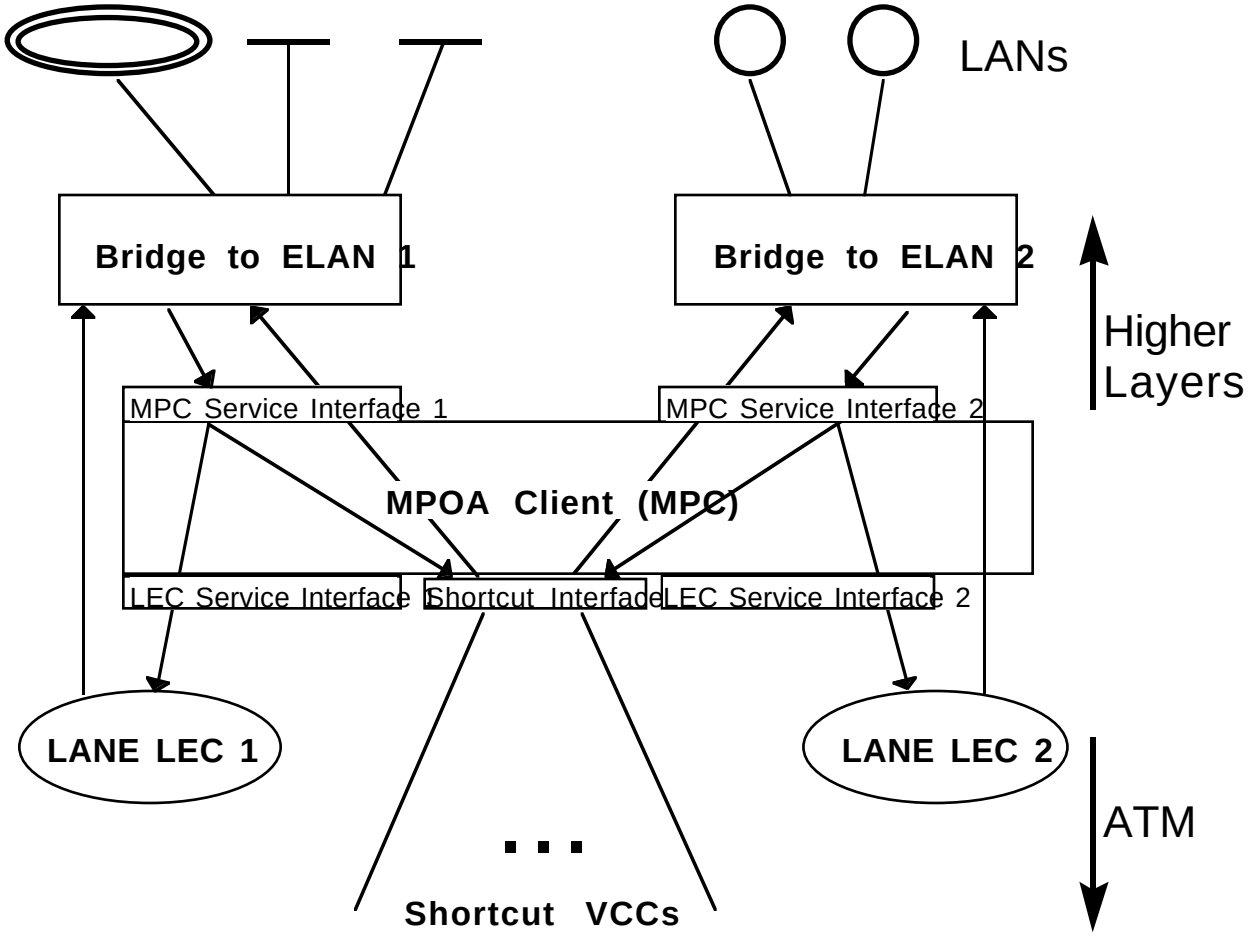


Figure 7 MPOA Edge Device MPC Example

For each LEC on which an MPC is to be enabled, the MPC supplies the LEC with the MPC's control ATM address via the MPOA Device Type TLV. Each LEC so notified includes the MPC device type TLV in every LE_ARP response which indicates to the recipient that an MPC is serving to the responding LEC, and indicates the control ATM address of the MPC.

The MPC presents the same Service Interfaces to its higher layers as its associated LECs present to it, except that the MPOA Device Type TLV may be stripped from the information provided to higher layers. An MPC analyzes packets from the MPC Service Interfaces for flow classification, collects statistics, and redirects appropriate packets to shortcuts. Non-redirected packets are passed on to the LEC Service Interface corresponding to the MPC Service Interface over which the packet was received. Packets received from a LEC Service Interface are passed transparently up to the corresponding MPC Service Interface. The difference between a LANE-capable bridge and an MPOA edge device lies in the MPC. Both LANE and bridging are outside the scope of MPOA. The detailed behavioral diagram of the MPC is shown in Figure 8. Note that the MPC sees only:

1. Packets sent by the higher layers of the MPC Service Interface destined for a LEC (i.e. the inbound data flow);
2. Packets received on a Shortcut Service Interface and relayed to the higher layers as if they came from a LEC (i.e. the Outbound data flow);

Data packets received on a LEC Service Interface destined for the higher layers are passed without examination by the MPC.

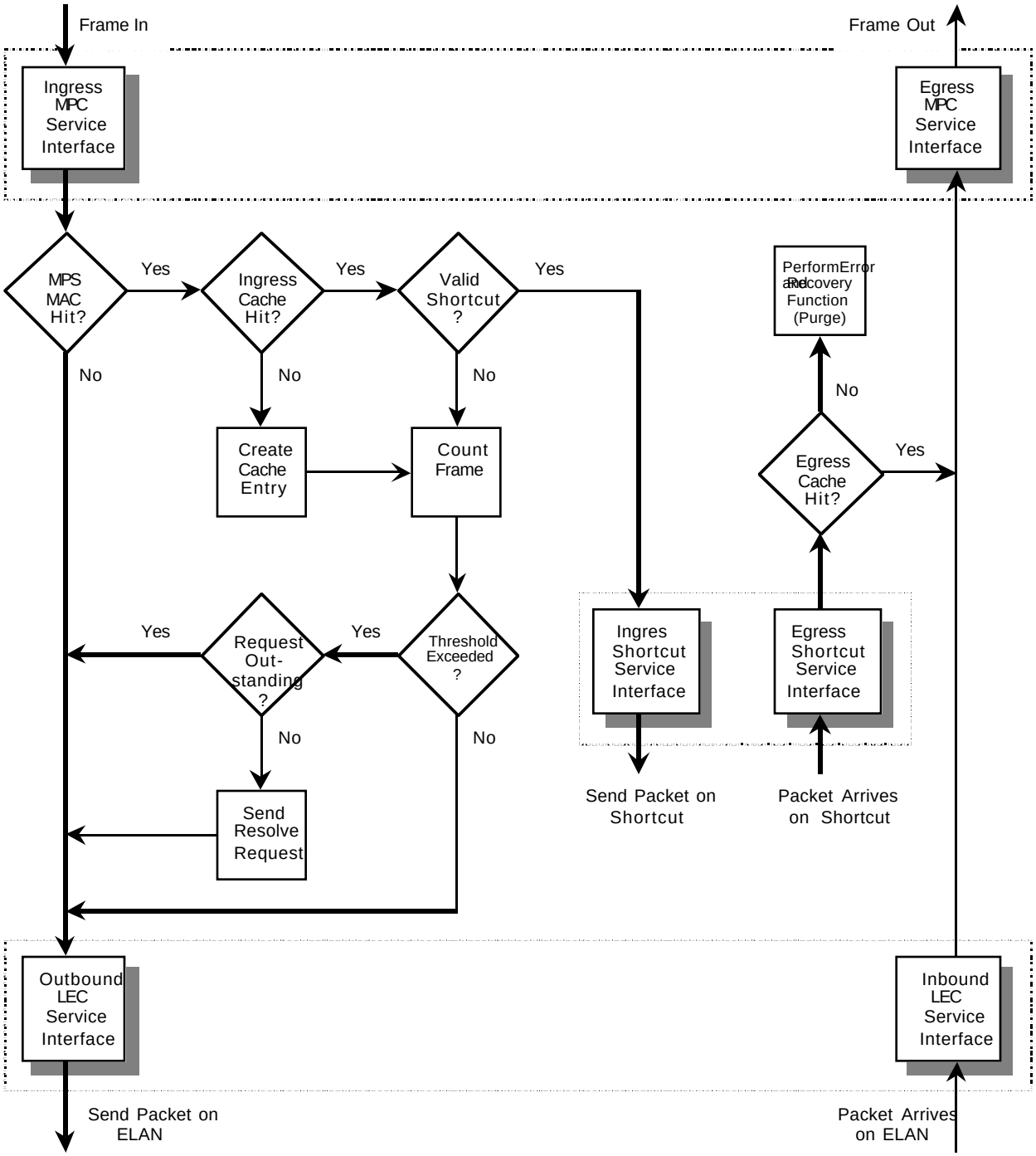


Figure 8 MPC Data Path Processing Logical Block Diagram

4.4.1 MPC Configuration

MPCs must be capable of obtaining configuration information from an LECS. To obtain this configuration information, the MPC sends a Configuration Request with a MPOA Device Type TLV identifying the device as an MPC to the LECS. The LECS uses the same mechanism when mapping Configuration Requests with MPOA Device Type TLVs to configuration information as it does for standard LEC requests. The LECS may use the MPOA Device TLV to return only MPC specific TLVs. MPC TLVs may be returned by the LECS in any

Configuration Response. MPC TLVs sent by the LECS override the default values of the corresponding MPC parameters. If no MPC TLVs are returned, the MPC must use default or locally configured values.

The MPC may be configured to bypass the configuration phase and use default or locally configured parameters. An MPC should retrieve configuration over only one of its LECs. MPC TLVs may be returned to a MPC's LEC during the LEC's initialization phase. The MPC may use this information without issuing another Configuration Request.

The TLV encodings of MPC configuration parameters are specified in Section 5.2.2.

4.4.2 Inbound Data Flow

All inbound packets (packets sent by the higher layers toward a LEC) are examined to see whether they have the destination MAC address of an MPS. The MAC addresses of MPSs on ELANs served by an MPC, and their associated control ATM addresses, are known to the MPC through the discovery mechanism described in Section 4.2.

If detection is enabled for the protocol in the packet, the MPC examines the internetwork layer destination address of the packet, and looks up that <MPS Control ATM Address, Internetwork Layer Address> tuple in its ingress cache.

It is important to note that, in a given MPC, more than one <LEC, MPS MAC address> tuple may map to the same MPS Control ATM address because both the MPS and MPC may be on more than one ELAN, and the MPS may have a different MAC address on each ELAN. In this instance, LEC refers to a particular LEC co-located with the MPC. Flow detection and requests, therefore, are on a per <MPS Control ATM Address, Destination Internetwork Layer Address> pair basis. An implication of this design is that when an MPS receives an MPOA resolution request, it does not know which of the LEC interfaces the packets related to the request would have arrived on; therefore, different inbound filters per interface cannot be supported with a single MPS Control ATM Address. To support per interface inbound filters, a router must use a different MPS Control ATM Address (and therefore a different MPS) per unique set of inbound filters.

The lookup process described above may be modeled as a two stage process:

1. <LEC, MAC> ⇒ MPS Control ATM Address (from LE_ARP cache).
2. <MPS Control ATM Address, Destination Internetwork Layer Address> ⇒ Cache Entry.

The contents of the ingress cache are shown in Table 2.

Table 2 Ingress Cache

Keys		**Contents**		
MPS Control ATM Address	Internetwork Layer Destination Address	Dest. ATM Address or VCC	Encapsulation Information	Other information needed for control (e.g. Flow Count and Holding Time)

If the <MPS Control ATM Address, Internetwork Layer Address> tuple is not found in the ingress cache, a new ingress cache entry is created. The ATM address/VCC field is invalidated, and the "count" field is set to 1 to count the frame. The frame is then sent on to the LEC for output to the ELAN.

If the <MPS Control ATM Address, Internetwork Layer Address> tuple is found, but the ATM address/VCC field does not specify an operational VCC, then the packet is counted in the count field. The frame is then sent on to the LEC for output to the ELAN.

By the time the count for a given <MPS Control ATM Address, Internetwork Layer Address> tuple exceeds the configured threshold for number of packets (MPC-p1) sent within a configured time period (MPC-p2), then the MPC is responsible for sending an MPOA Resolution Request to the MPS to which the packet's MAC destination address is associated, requesting a shortcut. The MPOA Resolution Request, specified in Section 5.3.2.4.6, is very similar to the NHRP Resolution Request. The main difference is that the source protocol address in the MPOA Resolution

Request may be NULL, since the MPC may not have an internetwork-layer address. The MPOA Resolution Request also must include an MPOA Tag TLV, should include an MPOA Service Category extension, and may include additional extensions as specified in [NHRP]. Note that during the period of time between when the threshold has been exceeded and a shortcut has been established, the MPC must limit the number of MPOA Resolution Requests it sends for the destination as described in Section 4.4.6.1.4.

If the <MPS Control ATM Address, Internetwork Layer Address> tuple is found in the ingress cache, and the ATM address/VCC field specifies an operational shortcut, then the packet's DLL encapsulation is stripped, the packet is encapsulated in the appropriate internetwork layer encapsulation (as defined in Section 5.1), and the packet is sent over the specified shortcut. The ingress MPC may receive padded frames. The ingress MPC may strip padding, or forward the frame with padding included.

4.4.3 Outbound Data Flow

Before an egress MPC can forward any packets, it must have received an egress cache entry from an egress MPS in an MPOA Cache Imposition Request. When an MPS determines that it must impose an egress cache entry on an MPC, as described in Section 4.5.2.2, the MPS sends an MPOA Cache Imposition Request to the MPC. The egress MPC must send an MPOA Cache Imposition Reply for every MPOA Cache Imposition request. To formulate its reply, the MPC must determine if it has the resources necessary to maintain the cache entry and potentially receive a new VCC. If the MPOA Cache Imposition is an update of an existing egress cache entry, the resources are likely available. If the MPC cannot accept either the cache entry or the likely resulting VCC, it sets the appropriate error status and returns the MPOA Cache Imposition Reply to the MPS. If it can accept this cache entry, it constructs an MPOA Cache Imposition Reply message. The MPOA Cache Imposition message includes an MPC data ATM address and a success status. The MPOA Egress Cache Tag Extension may also be included in the reply if it was included in the Cache Imposition Request. If an MPOA Service Category extension is present in a Cache Imposition Request, the egress MPC must set the Service Category extension in its Cache Imposition Reply to indicate the Service Categories supported by the egress MPC. If the MPOA Original Error Code TLV was in the request, the egress MPC must set it to the value as is set in the Cache Imposition Reply.

For all packets received on a shortcut, the egress MPC searches its egress cache for a matching entry. A cache hit is defined as a match on two main keys (source/destination ATM address pair, internetwork layer destination address) and/or one optional key (tag). In the non-tagged case, the source/destination ATM address pair is used as a key because it is possible for packets for a given internetwork layer destination to be forwarded to different next hops based on where they came from (see Appendix V).

The specific cache structure and lookup order is implementation dependent. The contents of the egress cache are shown in Table 3.

Table 3 Egress Cache Without Tags

Keys		**Contents**		
Internetwork Layer Destination Address	Source/Dest. ATM Addresses	LEC	DLL header	Other information needed for control (e.g. Holding Time)

An egress MPC may optionally use a tag for an egress cache entry. The tag may be used as the entire key, or as part of the key, to locate the relevant egress cache entry for a packet received on a shortcut VCC. The ingress MPC treats the tag as an opaque data string to be prepended to outgoing data frames.

If there are multiple egress cache entries that use the same source ATM address and internetwork layer destination address, but different DLL headers, then by using a different tag value for each one a unique key can be determined. An alternative method is to use a different destination ATM address, which also results in a unique key. An egress MPC may choose to associate a distinct tag value with every egress cache entry. In this case the egress cache entry can be determined by using the tag field alone.

Because an explicit indication of the internetworking protocol is not carried in packets using the tagged encapsulation on shortcut VCCs, an egress MPC must be able to determine the protocol from the egress cache entry retrieved

using the tag. The set of tag values used for different protocol address families, for a given source/destination ATM address pair, must therefore be distinct.

Table 4 Egress Cache With Tags

Keys			Contents		
Internetwork Layer Destination Address	Source/Dest. ATM Addresses	Tag	LEC	DLL header	Other information needed for control (e.g. Holding Time)

If an entry matching a packet received on a shortcut is not found in the egress cache, the packet is discarded, the error is counted, and the egress MPC initiates the Data Plane Purge process described in Section 4.7.2.3. If there is a cache hit, but the MAC destination address (from the DLL header) is not present in the LEC's C6, C8, C27, C30 variables, the egress MPC may continue to forward packets to the bridge as if they came from the LEC interface for up to 30 seconds to allow normal bridge flooding and learning procedures to occur. If the condition does not change within the 30 seconds, the egress MPC must invalidate the cache entry and send an MPOA Egress Cache Purge Request (See Section 4.7.1.6) to the MPS that imposed that egress cache entry.

If the VCC and internetwork layer destination address are in the cache, and the indicated LEC is fully operational, then the DLL header in the egress cache is attached to the internetwork layer packet, and the resultant frame is passed to the MPC service interface as if it arrived from the LEC. This DLL header comprises all octets preceding the start of the internetwork layer packet. The egress MPC may receive both padded and unpadded internetwork layer PDUs; therefore, the egress MPC may have to add padding to comply with the minimum frame size of the egress ELAN type.

The DLL header for the 802.3 format requires a length field indicating the length of all fields, starting with the LLC field, and including the IP packet. Therefore, the egress cache DLL header type is not transparent to the egress MPC. The egress MPC must parse the egress cache DLL header at least once, to determine that the length field is present, and must insert the correct value for the length of each outbound packet. Similarly, for inbound packets, the MPC must parse the frame to find the IP packet and check the validity of the length field, and then must perform the transformation to the shortcut format.

4.4.4 Cache Management

The ingress and egress caches are completely separate. Creation, deletion, or alteration of an entry in one cache does not imply any consequences for the other cache.

4.4.4.1 Ingress Cache Entry Creation and Management

An ingress MPC examines all packets destined for MAC addresses that belong to MPSs. When it detects a packet destined for an internetwork layer destination for which it does not already have a cache entry, it creates a new ingress cache entry for that internetwork layer destination. When the MPC detects a flow to a given internetwork layer destination, it sends an MPOA Resolution Request. When an MPOA Resolution Reply is received, the internetwork layer destination address, destination ATM address, source holding time, and MPOA Egress Cache Tag Extension (if present) are used to complete the ingress cache entry. If a reply is not received, the MPC should use the retry procedure defined in Section 4.2.3.

Any existing VCC may be used for data forwarding if its source and destination ATM addresses match those in the MPOA Resolution Reply, and the VCC signaling parameters are suitable. Otherwise, the ingress MPC must signal the creation of a new VCC before the shortcut can be used.

Ingress cache entries are aged using the source holding time from the latest MPOA Resolution Reply received for the corresponding internetwork layer destination address. Ingress cache entries may be withdrawn by the ingress MPS or deleted by the ingress MPC at any time for local reasons.

When an ingress MPC receives an NHRP Purge Request it must stop using the shortcut for packets destined to the specified internetwork layer destination address. It may issue a new MPOA Resolution Request immediately, or it may wait and use local information to determine when to query again.

To prevent active cache entries from aging out, ingress MPCs should issue new MPOA resolution requests to refresh these active cache entries at some time prior to the expiration of the holding time. For example, an MPC may choose to refresh active cache entries by sending a new resolution request after two thirds of the holding time has expired.

4.4.4.2 Egress Cache Entry Creation and Management

When an MPS determines that it must impose an egress cache entry on an MPC, the MPS sends an MPOA Cache Imposition Request to the MPC. The MPC uses the cache ID (in the MPOA DLL Header Extension) and the egress MPS control ATM address as the key to find and/or create an egress cache entry.

Egress cache entries are created with a holding time provided by the egress MPS. The entry must not be used beyond the egress holding time. If an egress cache imposition with a non-zero holding time is received for an existing cache entry, the egress MPC must update the egress cache entry. If an egress cache imposition with a zero holding time is received for an existing cache entry, the egress MPC must stop using the entry.

An egress MPC may find that it must discard packets received over a shortcut because the egress cache entry has become invalidated. The details of why a LEC no longer serves a LAN destination, or why an MPC views an egress cache entry as invalidated is a matter local to the MPC. An example would be an edge device that incorporates a bridge running the 802.1D spanning tree protocol, that finds that a packet received over a shortcut is due to be sent over a port that is not in the forwarding state, or is due to be sent back out the LEC port associated with the shortcut that the packet arrived on. This situation could occur due to a bridge topology change. Such a change might result in an edge device no longer being the correct edge device for a given internetwork layer destination address.

The definitions of the configuration and run-time variables controlling a LEC provide the means for specifying these conditions exactly. The LAN destinations served by a LEC are the union of the LAN destinations in the LEC's variables Local Unicast MAC Address(es) C6, Local Route Descriptor(s) C8, Remote Unicast MAC Address(es) C27, and Remote Route Descriptor(s) C30. Whenever a packet is received over a shortcut, the egress cache entry for that packet specifies the receiving LEC, and supplies the DLL encapsulation . If this destination MAC Address or route descriptor is not present in the LEC's C6, C8, C27, and C30 variables, then the egress cache entry is invalid and the MPC should initiate a purge.

Note that this follows normal LANE usage for answering LE_ARP_REQUESTs. No egress cache entry can be created if the destination MAC Address or destination Route Descriptor is not in the four LEC variables listed above, because the LEC would not answer an LE_ARP_REQUEST for that LAN destination. If the LEC would not answer an LE_ARP_REQUEST for the LAN destination, the router would not send that packet to that edge device via LANE, and hence the MPC has no business forwarding the packet.

If an egress MPC detects that an egress cache entry is invalid, it must inform the MPS that imposed the egress cache entry as described in Section 4.7.1.6. If the MPC has lost contact with the MPS, it should initiate a Data Plane Purge as described in Section 4.7.2.2.

4.4.5 LAN-to-LAN Flows Within the Same MPOA Device

An MPOA device must be able to handle the case where it is both the ingress MPC and the egress MPC for a given data flow. A simple implementation of this specification would cause the device to set up a VCC to itself. Although this simple implementation will work correctly, a more efficient implementation may build an internal shortcut and bypass the ATM network.

4.4.6 Control Information in MPC Caches

4.4.6.1 Ingress Cache

The control information required for maintaining ingress cache entries can be further divided into sub groups based on the function they are serving. The main control functions are State, Connection, Aging, Retry, Usage and Purge.

4.4.6.1.1 State Information

The following is a list of information that is needed to maintain the state of an ingress cache entry. The shortcut state information is separated from the VCC state information.

- Shortcut Entry State - Indicates if this is entry is in Resolving/Resolved/Invalid states.
- Shortcut VCC State - Indicates if the shortcut VCC is Open/Closed.
- Ingress MPS Control ATM Address - This is the address of the ingress MPS to which the MPOA Resolution Request is sent.
- Last NHRP CIE Code – This is the last CIE code received in a MPOA Resolution Reply. Tracking the last error aids in debugging.
- Last Q.2931 Cause Value – This is the last cause value received in a Q.2931 Release. Tracking the last error aids in debugging.

4.4.6.1.2 Connection Information

The packet forwarding function generally uses the VPI/VCI information, but there are other pieces of information that need to be maintained.

- ATM Address of Egress-MPC - This is the address used as the Called Party Address while setting up a shortcut VCC.
- Service Category - If an ATM Service Category Extension was received in a MPOA Resolution Reply, it should be saved to be used while setting up shortcut VCCs.

4.4.6.1.3 Aging Information

The MPOA Resolution Response returns a Holding Time that is the time for which the information returned in the response is valid. This information is used to age the entry once it is Resolved.

This field can be implemented in different ways; for example, it could be counted down to 0 based on a periodic timer or it could be set to the time at which this entry should be removed.

4.4.6.1.4 MPOA Resolution Request Retry

MPOA Resolution Requests may be retried if a response is not received within a reasonable amount of time. These retries must use the MPOA retry mechanism described in 4.2.3. In addition, if the request fails, a new resolution request for the same flow must not be issued for the duration of the Hold Down Time (MPC-p6).

The following information pertains to retries of MPOA Resolution Requests:

- Request Id - The Request Id that was used in an outstanding request, this should be reused for subsequent retries.
- RetryTime - This variable tracks the number of seconds that an MPC must wait before retrying a resolution request.

While attempting to resolve an address, an MPC may decided at any time that a shortcut is no longer needed and terminate the retry procedure.

4.4.6.1.5 Usage

There needs to be a count of the usage of a particular ingress cache entry for forwarding packets. This count is used primarily to do MPOA flow detection, but can also be useful for management and for maintaining lists for recycling cache entries when the system runs out of memory resources.

- Packets Forwarded - Number of packets that were forwarded using this cache entry.

4.4.6.2 Egress Cache

The control information contents of an egress cache entry can also be grouped based on the functions they serve. The different functions are State, Connection, Aging, Usage and Purge.

4.4.6.2.1 State Information

The state information consists of fields that are used to maintain the state of a shortcut. The state information of the entry is kept distinct from the actual VCC state.

- Shortcut Entry State - This field can be Resolved/Purge/Invalid. Packets received over shortcut VCCs are forwarded only when the entry is in the Resolved state
- Egress MPS Control ATM Address - This field is used to identify the MPS that imposed this cache entry. It is used for subsequent communication with the egress MPS for sending Egress Cache Purges.
- Cache Id - This information is used as a key in combination with the previous field to identify a unique egress cache entry for cache updates from the egress MPS.

4.4.6.2.2 Connection Information

The information related to shortcut VCCs are stored in these fields.

- Ingress MPC/NHC data ATM Address - The ATM address of the ingress MPC that issued the MPOA Resolution request for this entry. This address will be the Calling Party Address of an incoming shortcut VCC Setup Request. This may be used to verify the identity of the sender of packets over the shortcut VCC.

4.4.6.2.3 Aging Information

Every egress cache entry has a holding time associated with it that was provided in the MPOA Imposition Request message. This time is used to keep an entry in the Resolved State

4.4.6.2.4 Usage Information

It is useful for the management of the device to keep count of the number of packets that were received and processed using an egress cache entry. In addition this count would be useful to build a list of entries that need to be purged in case the system is running out of memory resources.

- Packets Received - This counter is incremented each time a packet is received on a shortcut VCC and this egress cache entry is used for encapsulating the packet before passing it to the outbound MPC Service Interface.

4.4.6.2.5 Purge Information

There are different reasons for purging entry/entries in an egress cache. The contents of the purge message vary depending on the purge reason.

- Egress MPS Control ATM Address - This information is needed if the MPC fails to receive an MPOA Keep-Alive message from the egress MPS within its life time. The egress MPC needs to inform the ingress MPC to purge all ingress cache entries that were imposed on the egress MPC by the failed egress MPS.

4.5 Detailed MPS Behavior

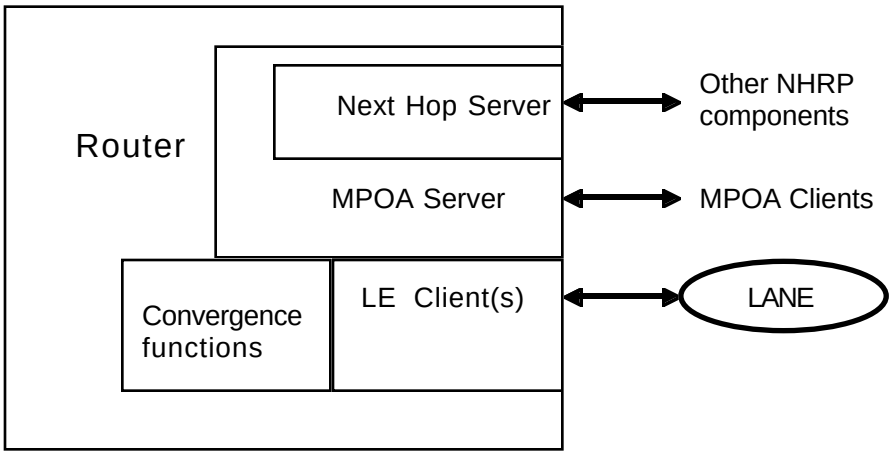


Figure 9 Router MPS Example

The MPS is a component of a router, as illustrated in Figure 9. An MPS is only useful in a router that has an NHS and interfaces to one or more LECs. The data and control path from the router through the LEC(s) to LANE is unaltered by MPOA. The MPS does, however, interact with the router, its LEC(s), the NHS, and other MPOA components. A LEC is associated with a single MPS.

The router engages in the operation of traditional routing protocols. One or more of the router's interfaces must use LANE.

The MPS must be aware of the router's configuration and forwarding tables to the extent of knowing (or being able to ask the router) whether a given internetwork layer destination address should be forwarded to a LEC, and which one. For each LEC and MAC Address on which the MPS is to be enabled, the MPS supplies the LEC with the MPS's control ATM address. Each LEC so notified includes the MPOA Device Type TLV in every LE_REGISTER_REQUEST, LE_ARP_REQUEST, and LE_ARP_RESPONSE response that it sends which to indicates to the recipient that an MPS is connected to the responding LEC and MAC Address, and indicates the control ATM address of the MPS.

Having advertised its control ATM address via LE_ARPLANE control messages, an MPS may receive MPOA Resolution Requests from MPCs. In addition, in its capacity as an NHS, the MPS/NHS may receive NHRP queries from NHCs or peer NHSs. The MPS/NHS handles both types of queries.

If the routing and subnet convergence information available to the MPS/NHS indicates that the next hop is an directly connected MPC, then the Resolution Request is passed on to that LEC's MPC as an MPOA Cache Imposition Request. Otherwise, the request should be treated as a standard NHRP Resolution Request and forwarded or answered as indicated in [NHRP].

The MPS must maintain the status of all ingress cache entries (positive MPOA Resolution Replies) and egress cache entries (positive MPOA Cache Imposition Replies) that it has given to its MPCs. The MPS will generate the reply, and record the fact of the reply. If the information is later invalidated, a notification will go to the source of the Resolution Request. A destination may become invalid either because the actual host moved/expired, or due to a routing change.

4.5.1 MPS Configuration

MPSs must be capable of obtaining configuration information from an LECS. To obtain this configuration information, the MPS sends a Configuration Request with a MPOA Device Type TLV identifying the device as an MPS to the LECS. The LECS uses the same mechanism when mapping Configuration Requests with MPOA Device Type TLVs to configuration information as it does for standard LEC requests. The LECS may use the MPOA Device Type TLV to return only MPS specific TLVs. MPS TLVs may be returned by the LECS in any

Configuration Response. MPS TLVs sent by the LECS override the default values of the corresponding MPS parameters. If no MPS TLVs are returned, the MPS must use default or locally configured values.

The MPS may be configured to bypass the configuration phase and use default or locally configured parameters. An MPS should retrieve configuration over only one of its LECs. MPS TLVs may be returned to a MPS's LEC during the LEC's initialization phase. The MPS may use this information without issuing another Configuration Request.

The TLV encodings of MPS configuration parameters information are specified in Section 5.2.1.

4.5.2 MPOA Resolution And NHRP Resolution

Ingress MPC-initiated MPOA Resolution includes a request phase and a reply phase, as shown in Figure 10. The role of the MPS in the Resolution process can be best described as that of a translator. An ingress MPC sends an MPOA Resolution Request to the appropriate ingress MPS. This ingress MPS translates the MPOA Resolution Request to an NHRP Resolution Request and forwards the Request on the routed path to the Internetwork-layer destination address (via its co-located NHS). When the NHRP Resolution Request arrives at the appropriate egress MPS, the egress MPS translates the NHRP Resolution Request to an MPOA Cache Imposition Request and sends it to the appropriate egress MPC. The egress MPC responds to the Cache Imposition Request by returning an MPOA Cache Imposition Reply to the egress MPS. The egress MPS then translates the MPOA Cache Imposition Reply to an NHRP Resolution Reply and forwards the reply on the routed path to the ingress MPS address in the NHRP Resolution Request (via its co-located NHS). When the ingress MPS receives the NHRP Resolution Reply, it translates the Reply to an MPOA Resolution Reply and returns it to the requesting ingress MPC. At the end of this process, the ingress MPC is prepared to establish and start using an MPOA shortcut and the egress MPC is prepared to receive data over the shortcut.

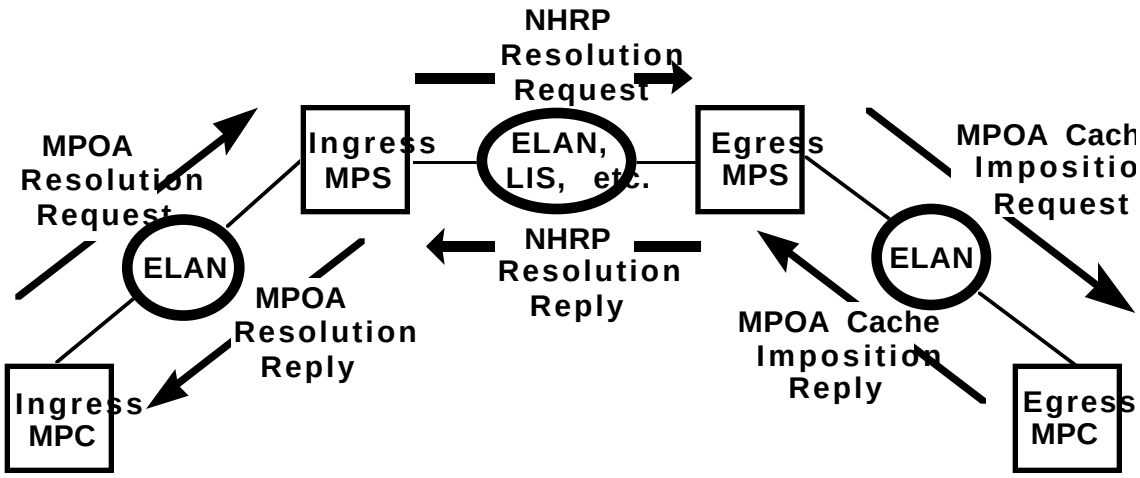


Figure 10 MPOA Resolution Process

MPOA Resolution Requests and MPOA Resolution Replies are identical in format to corresponding NHRP Resolution Requests and NHRP Resolution Replies except that a different Packet Type (ar$op.type) is used, as specified in Section 5.3.2.1. Distinction is required because MPCs are assumed to be associated with edge devices (i.e. bridges to LANs). This distinction results in protocol behavioral concerns not present with NHCs. Specifically, since the MPC does not necessarily have an internetwork layer address, the responding MPS/NHS may not be able to deliver the reply to the ingress MPC. Consequently, MPOA Resolution Requests are re-originated as NHRP Resolution Requests by the ingress MPS.

Re-origination ensures that corresponding NHRP Resolution Replies return to their point of origin. With MPS re-origination, MPOA models MPC-MPS relationships as distributed edge-routing: MPCs (as ATM addressed entities) are distributed forwarders associated with internetwork layer addressed (router co-resident) NHCs. Re-originated NHRP Resolution Requests contain the Source NBMA Address of the ingress MPC and the Source Protocol Address of the

router co-resident with the ingress MPS. The original Source Protocol Address (if any) and Request ID are retained by the ingress MPS to be re-inserted in later conversion of NHRP Resolution Reply to MPOA Resolution Reply.

It is important that the association of MPC data ATM address and router (MPS) protocol address is not "learned" by down-stream Next Hop Servers; in particular, if the authoritative responder is itself a non-MPOA NHS, it may misdeliver protocol messages (e.g. the NHRP Resolution Reply) to the MPC. This learning by NHSs is prevented by setting the S and D bits in the NHRP Flags field to 0 in NHRP Resolution Requests and Replies respectively.

The translations required by the MPS are explained in the following subsections.

An MPS has both ingress functions and egress functions. It is possible for an MPS to serve as both the ingress MPS and egress MPS for establishing and maintaining a particular shortcut. In this case, the resolution process is still modeled in this specification as an ingress MPS exchanging NHRP messages with an egress MPS through an NHS. A strict implementation of this model would result in a double translation; however, an actual implementation is free to optimize this case as appropriate.

4.5.2.1 Translating MPOA Resolution Requests to NHRP Resolution Requests

When an ingress MPS re-originates an MPOA resolution request as an NHRP resolution request, it maps the tuple <source Control ATM address, request id> to a unique request id for the re-originated NHRP resolution request. This tuple is maintained until an NHRP Resolution Reply is received or the MPS Give Up Time (MPS-p6) has expired. When an ingress MPS receives an NHRP resolution reply, it must convert it to an MPOA resolution reply and forward it to the requesting MPC. The Source Control ATM address and destination protocol address are retained for the Holding Time specified in the NHRP resolution reply so that the ingress MPS can correctly direct NHRP Purges.

The MPS must forward the reply to the MPC at the source ATM address of the VCC over which the MPOA resolution request was received (i.e. the MPC control ATM address). The MPS must NOT forward the reply to the source ATM address contained in the NHRP resolution reply (i.e. the MPC data ATM address). This behavior is intended to support MPC's which use distinct ATM addresses for control and data.

4.5.2.2 Translating NHRP Resolution Requests to MPOA Cache Imposition Requests

Prior to responding to an NHRP Resolution Request for an MPC, the MPS must impose an egress cache entry in the egress MPC by sending an MPOA Cache Imposition Request and receiving an MPOA Cache Imposition Reply.

When an MPS receives an NHRP Resolution Request from its co-resident NHS, it checks to see if the router forwarding tables direct that internetwork layer destination address to one of the LECs known to the MPS. If so, the MPS communicates with the appropriate router convergence functions (e.g. IP ARP) to determine the DLL header for frames sent through the LEC to that destination. The MPS must then check with the LEC to see whether the LAN destination used to reach that internetwork layer destination is served by an MPC. This information, along with the ATM address of the MPC, is passed via LANE LE_ARP control frames in the Device Type TLV, and is returned by the LEC to the inquiring MPS. Once this information is obtained, the MPS converts the NHRP Resolution Request to an MPOA Cache Imposition Request.

The egress MPS does this conversion by copying all fields in the Fixed and Common Header except as follows:

- ar$op.type is set to 0x80;
- Flags field is unused and must be set to zero;
- A new Request ID is generated.
- The Holding Time is set to at least twice the Holding Time the MPS will set in the corresponding NHRP Resolution Reply (MPS-p7 or as determined by local information).

In addition, message-specific portions and MPOA Extensions specified in Section 5.3.5 must be initialized. Included in this initialization is building and affixing to the message the MPOA DLL Header Extension . All NHRP Extensions included in the NHRP Resolution Request must be included in the MPOA Cache Imposition Request as

well. The MPS must retain sufficient information from the original NHRP Resolution Request to allow subsequent mapping of MPOA Cache Imposition Reply to NHRP Resolution Reply.

4.5.2.3 Translating MPOA Cache Imposition Replies to NHRP Resolution Replies

If the MPS receives a negative Cache Imposition Reply (NAK), it is the co-located NHS's responsibility to decide whether to accept the shortcut itself and return one of its own ATM addresses, or return a negative reply (NAK).

If a successful reply is received by the MPS, state for detecting routing changes is saved, and the reply passed to the original source of the Resolution Request. The MPS converts successful MPOA Cache Imposition Replies to NHRP Resolution Replies by copying all fields in the Fixed and Common Headers and CIE except as follows:

- ar$op.type is set to 2;
- Flags field is set as specified in [NHRP];
- Request ID is restored;
- Destination Protocol Address in the CIE is set to the egress MPS protocol address;
- Holding Time is restored to the value determined (i.e. less than or equal to half the holding time provided to the E-MPC).

Remaining NHRP Resolution Reply specific fields are filled in as specified in [NHRP].

If the NHRP Reply is generated with an MPC ATM address, the D (Destination Stability) bit must be zero to disable intermediate caching of the resolution.

The egress MPS must maintain state relative to all valid unexpired MPOA Cache Imposition Requests so that it may respond appropriately if the routing topology changes. If the cache imposition is successful, the egress MPS must maintain the mapping of internetwork layer address to DLL header and ATM address for the duration of the holding time it provides in the NHRP Resolution Reply as it would if it were forwarding the frames itself. In the case of IP, for example, the MPS must maintain its IP ARP cache entry in accordance with its locally configured ARP time-out parameter. If the holding time used in the NHRP Resolution Reply is greater than the IP ARP time-out, the MPS must re-verify the ARP when its time-out expires for the duration of the Holding Time. If a change is detected, the MPS must initiate the appropriate purge procedures.

4.5.2.4 Translating NHRP Resolution Replies to MPOA Resolution Replies

When the ingress MPS receives an NHRP Resolution Reply, the MPS converts this NHRP Resolution Reply to an MPOA Resolution Reply as follows. The MPS constructs an MPOA Resolution Reply containing the request ID and Source Protocol Address copied from the corresponding MPOA Resolution Request. It removes any TLVs it inserted. ar$op.type is set to 0x87. All other fields are copied from the NHRP Resolution Reply.

When an I-MPS receives an MPOA resolution request for which it is the logical next hop, it has two choices of what to do:

1. NAK the request and force the packets to be bridged via LANE.
2. Provide a reply using its own data ATM address.

Note that this will be a likely case when the MPS is in a router that is also the gateway to all off-campus destinations.

4.5.2.5 MPS to MPS NHRP

During the process of forwarding NHRP messages, an MPS/NHS may discover that the next hop is another MPS on a common ELAN as described in Section 4.2.

When an NHS co-resident with an MPS determines that a received NHRP message is to be forwarded to another NHS that is also co-resident with an MPS, the request shall be forwarded to the MPOA Control address of the peer NHS/MPS. The NHRP message shall be forwarded using AAL5 LLC/SNAP encapsulation as described in [NHRP].

When an MPS determines that the target of an NHRP Resolution Request is another MPS, rather than an MPC, it must forward the NHRP Resolution Request to that MPS as described above. This may occur when multiple MPSs share a common ELAN, and one MPS receives an NHRP Resolution Request for an internetwork layer address belonging to another MPSs on that ELAN.

4.6 Keep-Alive Protocol

MPCs need to know that MPSs that have supplied cache entries are alive and able to maintain those cache entries. As such, the MPS is required to periodically transmit an MPOA Keep-Alive message to all MPCs for which it has supplied and is maintaining ingress or egress cache entries. These must be sent every MPS-p1 seconds (subject to jitter). The MPOA Keep-Alive may be sent over any LLC/SNAP VCC between the MPS and the MPC. Specifically, it may be sent over a point-to-multipoint VCC, including one established specifically for the purpose of transmitting the Keep-Alive.

The Keep-Alive message contains the control ATM address of the MPS, a Keep-Alive Lifetime value, and a sequence number. The Control ATM address of the MPS is used to correlate cache entries with a particular MPS. The Keep-Alive Lifetime, specified as MPS-p2, is the length of time that a Keep-Alive message is to be considered valid. If a Keep-Alive message is not received within Keep-Alive Lifetime seconds (specified in the previous Keep-Alive message), the MPC must consider the MPS to have failed. If subsequent Keep-Alive messages received by an MPC do not have sequence numbers that increase in value, the MPC must assume that the MPS has rebooted and, therefore, also has failed.

If an MPC detects that an MPS has failed, it must invalidate all cache entries provided by that MPS. Note that there is no requirement that a VCC (either point to point or point to multipoint) be maintained between an MPS and MPC. In practice, the keep-alive traffic itself will typically prevent the VCC from idling out, but this depends on the relative time-out values. In particular no inference about the state of an MPS should be drawn by an MPC if a VCC is released over which keep-alive messages were being received. The inference that an MPS is down should only be drawn after the Keep-Alive Lifetime has expired.

4.7 Cache Maintenance

Ingress and egress cache entries are created through the MPOA resolution process. Once created, these cache entries must be updated or removed as appropriate. This section describes the mechanisms provided to perform this cache maintenance.

The following purge mechanisms are used by MPOA components:

1. MPOA Cache Imposition Request from egress MPS to egress MPC to either refresh or purge a cache entry
2. MPC-Initiated Egress Cache Purge from egress MPC to egress MPS
3. NHRP Purge Request from ingress MPS to ingress MPC
4. NHRP Purge Request sent on the data plane from egress MPC to ingress MPC

4.7.1 Egress Cache Maintenance

An MPS must maintain state for all the MPOA and NHRP Resolution Replies and successful MPOA Cache Imposition Requests that it sources for the duration of the holding time it provides. The holding time provided by the MPS is viewed as a contract in that the MPS guarantees that, for the duration of the holding time, if the information it gave to another party changes, it will send a notification (update or purge) to that party. The recipient of the information is then free to use the information provided by the MPS for the duration of the Holding Time (unless it detects a change). The items for which an MPS maintains state are called “active cache entries”.

From the perspective of an egress MPS, active cache entries are those for which it has performed a successful MPOA Cache Imposition Request and answered an NHRP Resolution Request.

4.7.1.1 Egress MPS Purges and Cache Updates

When an egress MPS detects a change for a destination internetwork layer address affecting one of its active egress cache entries, it must:

1. Send NHRP Purge Requests to the set of affected sources of relevant resolution requests, and
2. Send an MPOA Cache Imposition Request with a holding time of zero to the egress MPCs with affected egress cache entries.

Or, it must:

1. Send an MPOA Cache Imposition Request with an updated DLL header to the egress MPCs to update the affected egress cache entries. If a DLL header update is sent, and the corresponding MPOA Cache Imposition Reply contains any new information (e.g., a new tag, a new data ATM, or different TLVs), then the MPS must send a purge as described above because the ingress MPC cannot be updated with an unsolicited resolution response.

The reasons that a relevant change may occur include:

- Routing has changed such that:
    - The egress MPS/NHS is no longer the NHRP Authoritative Responder, so that a received NHRP Resolution Request would be forwarded to another NHS;
    - The next internetwork layer hop has changed so that a received NHRP Resolution Request would cause an MPOA Cache Imposition Request to be forwarded to different MPC;
    - The internetwork layer next hop has changed so that a received NHRP Resolution Request would cause an MPOA Cache Imposition Request to be forwarded to the same MPC with a different DLL header.
- Bridging over LANE has changed such that:
    - The egress LEC has changed so that the shortcut needs to go to a different MPC than previously given (generally detected by LE_ARP).
    - An Egress Cache Purge Request has been received from an egress MPC

To update an egress cache entry, the egress MPS sends an MPOA Cache Imposition Request with the same egress Cache ID that was used on the original Cache Imposition Request and a non-zero holding time.

When an egress MPC receives an MPOA Cache Imposition Request with a Cache ID matching an active egress cache entry received from the same MPS, it must replace the fields in the current egress cache entry with corresponding fields in the new MPOA Cache Imposition Request.

To purge an egress cache entry, an egress MPS sends an MPOA Cache Imposition Request with the same egress Cache ID that was used on the original Cache Imposition Request and a zero holding time. The egress MPS may purge all egress cache entries in an MPC for a given destination protocol address by including the protocol address in the CIE and omitting the MPOA DLL Header Extension (which would contain the cache ID).

To ensure that correct updates are made in either the case of an update or a purge, the egress MPS must send the MPOA Cache Imposition Request to the egress MPC using the same VCC (or a VCC originating from the same egress MPS control ATM address and terminating at the same MPC ATM address) as was used for the original cache imposition.

Egress cache updates must be sent reliably using the retry mechanism described in Section 4.3.

4.7.1.2 Egress MPC Invalidation of Imposed Cache Entries

An egress MPC must invalidate any imposed egress cache entry for which the holding time has expired.

An egress MPC must invalidate all egress cache entries that originated from an egress MPS with which the MPC has lost communication, as described in Section 4.6.

If an egress MPC receives a packet on a shortcut, and the corresponding egress cache entry specifies a MAC destination address or destination Route Descriptor that is no longer in any of the associated LEC's variables (Local Unicast MAC Address(es) C6, Local Route Descriptor(s) C8, Remote Unicast MAC Address(es) C27, and Remote

Route Descriptor(s) C30), it takes the following actions. The egress MPC may continue to forward packets to the bridge as if they came from the LEC interface for up to 30 seconds to allow normal bridge flooding and learning procedures to occur. If the condition does not change within the 30 seconds, the egress MPC must invalidate the cache entry and send an MPOA Egress Cache Purge Request (See Section 4.7.1.6) to the MPS that imposed that egress cache entry.

4.7.1.3 Invalidation of State Information Relative to Imposed Cache

An MPS must assume that all egress cache entries imposed by it to an egress MPC with which it has lost all communication may continue to be used until the Holding Time expires, or until it has sent egress MPS-initiated cache purges as described in section 4.7.1.1 and must not remove state information relative to these impositions. The MPS must expire this state information normally and may re-impose egress cache entries associated with remaining state information on restoration of the connection to the egress MPC.

4.7.1.4 Recovery From Receipt of Invalid Data Packets

An egress MPC that continues to receive data on a shortcut for which it does not have a valid egress cache entry must periodically send an MPOA Data Plane Purge to the ingress MPC as defined in Section 4.7.2.3. The frequency of these purges should not exceed one per second per source ATM address/destination internetwork layer address pair. This is required to recover from situations that may arise as a result of a lost cache imposition or incorrect shortcut usage by the remote end.

4.7.1.5 Egress Encapsulation

Cache impositions contain DLL encapsulation information as defined in an appropriate Annex to this document (e.g. - Annex A describes protocol specific encapsulation used for IP and IPX).

4.7.1.6 MPC-Initiated Egress Cache Purge

The MPC-Initiated Egress Cache Purge protocol provides the capability for an egress MPC to notify the egress MPS when it discovers an invalid egress cache entry. This notification allows the egress MPS to issue associated NHRP Purge Requests. The MPC-Initiated Egress Cache Purge Request is most likely used when either the bridge topology has changed and a destination is no longer behind the same edge device, or the destination has aged out of the bridge forwarding table for lack of communication.

Information to be included in an MPOA Egress Cache Purge Request is:

- Request ID
- Egress MPS Protocol Address
- Egress MPC Data ATM Address
- Destination Protocol Address (to purge)
- Destination Prefix Length
- Ingress MPC Data ATM Address (Optional)
- MPOA DLL Header Extension (Mandatory)
- no-reply flag

Upon receiving an MPOA Egress Cache Purge Request, the egress MPS must generate the appropriate NHRP Purge Request for the entry indicated in the Egress Cache Purge Request.

The no-reply flag (N-bit) is used to indicate whether the egress MPC wishes to receive an MPOA Egress Cache Purge Reply. If the no-reply flag is cleared, an MPOA Egress Cache Purge Reply is expected and the MPS must clear the no-reply field in the associated NHRP Purge Request. When the egress MPS receives the associated NHRP Purge Reply, it issues an MPOA Egress Cache Purge Reply to the egress MPC. In the Egress Cache Purge Reply, the egress MPS returns all information provided by the egress MPC in the request.

If the no-reply flag is set in the MPOA Egress Cache Purge Request, the egress MPC does not expect to get an MPOA Purge Reply. If an egress MPC does not request an MPOA Egress Cache Purge Reply, it is a local matter to the egress MPS/NHS whether to request an NHRP Purge Reply.

If the shortcut is between an ingress MPC and an egress MPC, the NHRP Purge Request is sent to the ingress MPS (identified by its internetwork layer address) that re-originated the NHRP Resolution Request after receiving the original MPOA Resolution Request from the ingress MPC. The ingress MPS then forwards the NHRP Purge Request to the ingress MPC. Note that multiple ingress cache entries may be invalidated as a result of a single MPOA Egress Cache Purge Request. This is because the scope of the NHRP Purge Request includes all entries covered by the source, destination and internetwork layer destination addresses in the NHRP Purge Request and is not restricted to the source and destination ATM addresses of the shortcut.

4.7.2 Ingress Cache Maintenance

4.7.2.1 MPOA Trigger

An MPC must be able to detect inbound data flows and establish shortcuts. In addition, an ingress MPS may detect inbound data flows and request that ingress MPCs establish shortcuts for them. A trigger mechanism is used such that the rest of the protocol remains consistent with the MPC-initiated mechanism.

In the event that an ingress MPS determines the need for a shortcut for an inbound data flow, the ingress MPS may trigger the appropriate ingress MPC into initiating an MPOA Resolution Request for that flow. This is done using an MPOA Trigger that describes the inbound data flow to be shortcut. The ingress MPC must create an ingress cache entry for the flow, if one does not already exist and if it has the resources to establish another shortcut, and must respond by initiating MPOA Resolution Requests for the target indicated in the MPOA Trigger.

An ingress MPS must use the MPOA retry procedure defined in Section 4.2.3 to control the sending of MPOA Trigger messages (note that the receipt of a corresponding MPOA Resolution Request by the triggering MPS is considered to be the reply for a given MPOA Trigger message

Information provided in an MPOA Trigger is:

- Ingress MPS Control ATM Address
- Destination Internetwork Layer Address and Address Prefix

The meaning and use of these information fields is given in the sections below.

Ingress MPS Control ATM Address

This address is required for an ingress MPC to build an ingress cache entry and identify the inbound datagrams that will be sent on the shortcut to be established as a result of an MPOA Resolution Request.

Destination Internetwork Layer Address and Address Prefix

The address to be used in the triggered MPOA Resolution Request. This address is also required for an ingress MPC to build an ingress cache entry and identify the inbound datagrams that will be sent on the shortcut to be established as a result of a successful receipt of a corresponding NHRP Resolution Reply.

4.7.2.2 Ingress MPSs and NHRP Purges

When an ingress MPS receives an NHRP purge request, it must send NHRP purge requests to all relevant MPC's for which it is maintaining ingress state for the purged destination address(es).

If the no-reply flag is clear in the received NHRP purge request (meaning an NHRP purge reply is requested), the ingress MPS must use the retry procedure, defined in Section 4.3, to ensure reliable delivery of NHRP purge requests to relevant MPCs. The ingress MPS may send an NHRP purge reply without waiting for NHRP purge replies from all MPC's.

4.7.2.3 Data Plane Purge Protocol

Under certain circumstances it is necessary to send an NHRP Purge Request on the data plane to tell the ingress MPC or NHC that ingress cache entries are no longer valid.

The different conditions under which an egress MPC is required to send NHRP Purge Requests over the shortcut are described below:

Egress MPS Dies: If an egress MPC fails to receive an MPOA Keep-Alive message from an MPS that has imposed egress cache entries within the MPOA Keep-Alive Lifetime (as specified in the last received MPOA Keep-Alive message) then it must send NHRP Purge Requests which invalidate all the cache entries imposed by the failed MPS, which are currently associated with a shortcut VCC. If there is no open VCC to the source ATM address as specified in an egress cache entry, it is not necessary to establish a VCC for the purpose of sending an NHRP Purge Request.

Egress Cache Miss: If an MPC receives a packet over a shortcut, but the egress cache lookup fails, the MPC must send an NHRP Purge Request over that shortcut to inform the ingress MPC to remove the appropriate ingress cache entry. In this case, the egress MPC does not know to which MPS to send an Egress Cache Purge request.

Note: An egress cache Miss can occur for several reasons:
1. Destination internetwork layer address not found.
2. Invalid tag.
3. IP/tag/source ATM address not consistent (e.g. tag hit, but wrong destination IP address or destination IP hit, but wrong source ATM).

The Data Plane Purge mechanism uses the NHRP Purge Request frame format as described in Section 5.3.11. When an ingress MPC receives the NHRP Purge Request on the shortcut, it must do the following in this order: authenticate the NHRP Purge Request if authentication was used, process any vendor private Extensions, process MPOA-specific Extensions, and purge the ingress cache as appropriate. In the case of conflicting information in Extensions, the previously specified order also specifies the priority for conflict resolution (e.g., do nothing if authentication fails.)

4.8 Connection Management

This section specifies the procedures for MPOA connection management in their entirety. These procedures are derived from, and are intended to be compatible with, those described in [RFC 1755], which describes procedures for establishing and clearing VCCs in a multiprotocol environment. In some cases text from [RFC 1755] is reproduced here in this specification either unmodified or slightly modified, without an explicit reference that the text has been taken from that source, but the work done by the authors of that RFC is hereby acknowledged.

4.8.1 Generic VCC Management Procedures

MPOA components must support the use of LLC/SNAP encapsulation for all PDUs. By default VCCs must be signaled to use LLC encapsulation. The negotiation of other encapsulations, such as the null encapsulation, is not precluded, but an MPOA component is not required to support any encapsulation other than LLC/SNAP.

Because the connection management procedures use VCCs with LLC encapsulation, many of the procedures are generic procedures that can be used by any protocol. The same VCC may be used to carry both MPOA control and data traffic. Also the same VCC may be used to carry both MPOA and non-MPOA traffic, as long as the non-MPOA traffic uses LLC encapsulation. Where the MPOA protocols require specific rules or default values these are explicitly indicated. Apart from that, the text can be read as applying to any protocol that uses LLC encapsulation in an MPOA-capable device.

Although LLC encapsulation allows the sharing of a VCC by multiple protocols, it does not require it. Stations can control the ATM addresses that they advertise for different protocols (using a different ATM selector for example),

forcing separation. Also, when a station is transmitting traffic, it may establish multiple VCCs between the same two endpoints, each carrying a different protocol, for example.

Communication between multiple local and remote protocol entities may use a single VCC if the local entities all are sharing an ATM address and the remote entities are all sharing an ATM address. Such sharing is facilitated by using LLC multiplexing on the VCC.

The MPOA specification imposes rules on the allocation of ATM addresses within an MPOA device and, as a result, on what entities may share a VCC. For example each MPC in a device must use a distinct ATM control address. Also, the ATM address assignment for LANE Data Direct VCCs is constrained as described in Section 4.2.3. The sharing of VCCs is thus always constrained by the overarching ATM address assignment rules. Within those constraints an implementation is free to allocate the same, or different, ATM addresses to different protocol entities. For example an MPC may choose to use one ATM address for MPOA shortcut VCCs, and another for LANE LLC Data Direct VCCs, or it may choose to use a single ATM address for both.

Note that there is no necessity to have one particular protocol or protocol suite as a primary owner of a VCC. Whether there is one protocol that "owns" a VCC and allows sharing, or a separate VCC management entity to which all protocols make requests as peers, is purely an implementation issue and as such is outside the scope of this specification. Note that because the VCCs are potentially shared, it is not possible to deduce status information about a particular protocol based on status information of a particular VCC. In particular, it is not possible to deduce that a protocol entity is operational just because a VCC has been established, as the process or task implementing that protocol could be non-operational.

4.8.2 Scope of MPOA VCCs

PDUs are sent between MPOA components and also between MPOA components and non-MPOA components, in particular between MPCs and NHCs that do not also include MPOA functionality. Note that packets transmitted between MPOA capable routers use NHRP between the co-located NHSs. It is possible to mix routers with MPSs and routers with NHSs in a network. An NHRP Resolution Request issued by an MPS may be answered by a router with an NHS; for example, if the router has directly attached Ethernet hosts, and is the egress router for those hosts, it may answer the Resolution Request.

An MPOA component must be capable of establishing, receiving and maintaining a VCC to any entity that conforms to the connection management procedures specified in this document, whether or not that entity is an MPOA component.

4.8.3 Initiating VCCs

VCCs are established when needed. If an MPOA component has a datagram to send and there is no existing VCC that it can use, or it chooses not to use an existing VCC, then it establishes a VCC to transfer the datagram. Both MPCs and MPSs may initiate calls. For example an MPC may initiate a call to transfer an MPOA Resolution Request, and an MPS may initiate a call to transfer an MPOA Trigger or NHRP Purge Request.

When an MPOA component has a datagram to send, it should first look to see if there is an existing VCC that it can use. There may be an existing VCC to the correct ATM address that it chooses not to use, for example, due to a mismatch in AAL5-SDU sizes or because the additional traffic could violate the Traffic Descriptor used when the VCC was first established. In some cases, an MPOA component may choose not to use an existing VCC for its own local purposes, such as to achieve protocol separation. If an MPOA components chooses not to use an existing VCC, then it must attempt to establish a new VCC.

4.8.4 Receiving Incoming VCCs

When an incoming VCC is established that indicates the use of LLC encapsulation, there is no information conveyed by the UNI signaling protocol about what protocols will subsequently be used over the VCC. For example, the originator may use the VCC for control or data traffic or both. Unless limited by resources, MPOA components should accept all incoming VCCs that indicate the use of LLC encapsulation.

Any form of security checking before incoming call acceptance (e.g. by determining if the calling ATM address belongs to a set of valid addresses) is outside the scope of this specification. An MPOA component must be prepared to receive incoming calls originated in the same ELAN or in a different ELAN, if it has any mechanism to differentiate between the two based on the calling party ATM address. Any such mechanism is outside the scope of this specification. An MPOA component must not make any assumptions at call setup time about the type of traffic (e.g. Control or Data) to be used on a VCC based on information as to whether the Calling Party is in the same or a different ELAN.

4.8.5 Support for Multiple VCCs

MPOA components must be able to support multiple VCCs between peer systems, without regard to which peer system initiated each VCC. When an incoming call is accepted, an MPOA component must be prepared to receive incoming PDUs on that VCC. It may also transmit PDUs on that VCC. It must not accept and immediately clear the incoming call, or ignore PDUs received on the VCC.

Allowing multiple VCCs is primarily intended for cases where the VCCs have different attributes, such as Traffic Descriptor, Quality of Service requested from the network, or AAL5-CPCS-SDU size. It is recognized that independently of these considerations two MPOA components may simultaneously initiate calls to each other, leading to duplicate identical VCCs. To avoid the wasted resources of unintentional duplicate VCCs, the following mechanism is defined to allow one VCC to be cleared due to inactivity, with all traffic carried on the other VCC.

When an MPOA component has a datagram to send, and it detects that there are more than one VCC that are capable of conveying the packet, and that the MPOA component does not wish to use more than one VCC, it should send the packet on the VCC initiated by the party that has the numerically lower ATM address. In this way both parties will use a single VCC, allowing the other VCC to time out due to inactivity. If an MPOA component switches over to use a VCC in this way, it may set the inactivity timer to a small value for the VCC that it does not intend to use, provided that it was the initiator of the VCC that it does not intend to use. If an MPOA component finds during this process that there are multiple VCCs initiated by the party with the lower value of ATM address, it may choose to use any one of them.

Specific protocols that use LLC encapsulation may impose more restrictive rules. In particular, a protocol may mandate that only one VCC be used between a pair of end stations to transfer traffic of that protocol between those two end stations. Such a per-protocol restriction does not affect the use of multiple VCCs by other protocols in the same box, and may coexist in a device that uses protocols that allow multiple VCCs.

4.8.6 Internetwork Layer-to-ATM Address Mapping

MPOA components are not required to support the InATMARP protocol as defined in [RFC 1577]. One reason for this is that MPCs are not required to have any internetwork layer addresses. An MPOA component is not permitted to learn Internetwork Layer Address-to-ATM Address mappings as a result of using the InATMARP protocol on VCCs.

An MPOA component must not learn Internetwork Layer-to-ATM Address mappings by learning from the Source Protocol and Source NBMA Address fields in an MPOA/NHRP Resolution Request. The only method used to learn Internetwork Layer Address-to-ATM Address mappings, apart from static configuration, is to issue an MPOA/NHRP Resolution Request and learn from the MPOA/NHRP Resolution Reply. One reason for this is that there may be asymmetrical routes in the network. Another is that the Source Protocol address in an NHRP Resolution Request may be that of an MPS, and the Source NBMA Address that of a physically separate MPC.

4.8.7 Establishment of Bi-directional Data Flow

When a VCC is established by an ingress MPC to an egress MPC, traffic in the reverse direction (egress MPC to ingress MPC) may use that same VCC, as described in 4.4.4.1, if the signalling parameters in that direction are suitable. If desired, an MPOA component can establish a dedicated unidirectional VCC by specifying a return PCR of zero.

4.8.8 VCC Termination

There is no requirement that a VCC be permanently maintained between two MPOA components. Either the Calling Party or the Called Party may clear the VCC. If a VCC is terminated by the remote party, a station should not immediately re-establish the VCC unless it has some PDUs to transfer.

In general, an MPOA component should allow VCCs to idle out based on inactivity. Additionally, an MPOA component may decide to release the least recently used VCC to free up resources for a new VCC, or as a reaction to certain error conditions, such as persistent protocol errors due to traffic on a certain VCC.

4.8.9 Use of UNI Signaling Information Elements

MPOA control and data PDUs may be transferred on any suitable VCC. Such a VCC may have previously been set up to transfer a PDU for a non-MPOA protocol, but that has not yet timed out. If an MPOA component wishes to setup a new VCC to transfer an MPOA PDU, then the following rules apply with regard to the encoding of UNI signaling information elements. It is not a requirement that an MPOA PDU be transferred over a VCC that was established in accordance with the these rules. It is a requirement that an MPOA component be capable of establishing a VCC according to these rules.

Shortcut data flows, as specified in this version of MPOA, are always carried over point-to-point VCCs. In general, Control PDUs will also be carried over point-to-point VCCs, but point-to-multipoint VCCs may also be used in specific cases. An MPOA component must be able to be added as the first or subsequent party of a point-to-multipoint call. The use of the signaling IEs is the same as that outlined here for point-to-point calls, subject to the constraints imposed by [UNI 3.1]. The main difference is that call characteristics, such as Traffic Descriptor, QoS, or AAL5 Parameters, can only be negotiated between the Calling Party and the first Called Party. The second and subsequent Called Parties cannot engage in any negotiation. An MPOA component is not required to be able to initiate point-to-multipoint calls for Control PDUs. Note that the intra-ELAN multicast and broadcast data flows are handled by the LANE BUS, and are transparent to MPOA. The use of shortcuts for multicast data flows, i.e. bypassing multicast routers, is not supported in this specification.

To enhance interoperability, an MPOA component must treat as equivalent the presence of a parameter set to a null or zero value, and the absence of that parameter. All unused and reserved fields must be set to zero on transmission and ignored on receipt. A received packet that has non-zero values in these fields must not be treated as an error.

4.8.9.1 Traffic Descriptor

All MPOA components must be capable of initiating and accepting VCCs with the UBR service category. The support of VCCs with other than the UBR service category is allowed, but not required.

For transferring control messages, an MPOA component should initiate a VCC with the UBR service category. If an MPOA component attempts to set up a VCC using anything other than the UBR service category, for the purposes of transferring control messages, and the VCC establishment fails as a result of either the network or the remote party being unable to support a non-UBR service category, the MPOA component must retry using the UBR service category.

For shortcuts, an MPOA component must be capable of initiating a VCC that proposes the UBR service category. It may propose any service category, but must be prepared to deal with either the network or the remote party rejecting the call due to being unable to support the proposed non-UBR service category. In this case the MPOA component should retry using the UBR service category.

The mechanisms by which an MPOA component decides to initiate a VCC that uses anything other than the UBR service category, are outside the scope of this specification. Different mechanisms are possible such as monitoring the data flow, or monitoring or participation in a resource reservation mechanism like RSVP [RSVP].

UNI 3.x Signaling does not provide for ATM Traffic Descriptor or Quality of Service negotiation. UNI 4.0 Signaling does provide for the negotiation of these parameters within an ATM service category, but does not permit the negotiation of the ATM Service Categories themselves. To determine what ATM Service Categories a target station supports , an MPOA component should use the ATM Service Category Extension in advance of VCC

establishment. This capability reduces the probability that a VCC will be rejected because the Service Category cannot be supported, with the result that the calling party has to try again.

When an ingress MPC sends an MPOA Resolution Request, it should add a single ATM Service Category Extension, as specified in Section 4.4.2, to identify the Service Categories it supports. If only UBR is supported the extension should still be added, and will take a zero value. When an egress MPC responds to an MPOA Cache Imposition Request with an MPOA Cache Imposition Reply, it should fill in the ATM Service Category Extension to identify the Service Categories it supports, if the extension was included in the request, as specified in Section 4.4.3.

When an ingress MPC knows, through receipt of an ATM Service Category Extension in an NHRP Resolution Reply, that the desired Service Category is supported on the target MPC, it may attempt to set up the shortcut with that Service Category. If the first attempt for the call setup does not succeed, the MPC may attempt with another Service Category that both ends support. The MPC may attempt to setup a shortcut with the UBR Service Category at any time.

The PCR used in the ATM Traffic Descriptor should be set to line rate. It may be set to less than line rate as a local configuration option. This may be useful if it is known that there are slower speed links in the network, and that traffic shaping to the slower speed by a device may reduce frame loss. As specified in section 3.6.2.4 of [UNI 3.1], for best effort traffic, a user need not conform to the signaled PCR, and the network may enforce a PCR different than the signaled PCR. It is recommended, however, that if a user signals a PCR less than line rate that it conform to the PCR.

The valid combinations of values of the Broadband Bearer Capability, Traffic Descriptor and QoS Class IEs are defined in Appendix F of [UNI 3.1].

Use

This IE must be included in a SETUP message.

Format

Field	Value
Forward Peak Cell Rate (CLP=0+1) ID	132
Forward Peak Cell Rate (CLP=0+1)	line rate
Backward Peak Cell Rate (CLP=0+1) ID	133
Backward Peak Cell Rate (CLP=0+1)	line rate
Best Effort Indication	190

UNI 3.0 considerations

In UNI 3.0, issues of traffic management were less well understood than in UNI 3.1. UNI 3.0 does not contain a guide to coordinating the use of the User Cell Rate IE (Traffic Descriptor in UNI 3.1), Broadband Capability IE, and QoS Parameter IE. It is recommended that the use of these IEs in UNI 3.0 should be the same as for UNI 3.1.

The value for the cause “User Cell Rate” is 51 in UNI 3.0, and 37 in UNI 3.1

UNI 4.0 considerations

An MPOA component must support the ability to set both the forward and backward frame discard bits. An MPOA component must use the frame discard capability. This capability may be used with any ATM service category. The use of frame discard increases throughput under network congestion. An MPOA component must not treat the reception of an ATM Traffic Descriptor IE which does not indicate frame discard as an error.

Traffic parameter negotiation using the Minimum Acceptable ATM Traffic Descriptor IE is an optional UNI 4.0 feature. If available an MPOA component should use this capability. This capability may be used with any ATM service category. The use of traffic parameter negotiation allows the calling and called parties to discover the smallest bandwidth limitation along the path of the connection. With this information a source may then take measures to reduce the possibility of network congestion. For example a source may then perform traffic shaping down to the smallest bandwidth, or may perform congestion control at a higher layer on an end to end basis. Even if the calling party is not capableof using the information about the smallest bandwidth the MinimumAcceptable ATM Traffic Descriptor IE should be used, as the called partymay be capable of doing so.

When using traffic parameter negotiation with the UBR service category the PCR should be set to link rate in the ATM Traffic Descriptor IE, and the PCR should be set to zero in the Minimum Acceptable ATM Traffic Descriptor IE. Note that the value of zero in the Minimum Acceptable ATM Traffic Descriptor IE will be treated by the network as meaning "unspecified", not that zero bandwidth is acceptable.

When progressing a call the network will adjust the ATM Traffic Descriptor IE to reflect the actual bandwidth available. The network may drop the Minimum Acceptable ATM Traffic Descriptor IE from the SETUP message, if it and ATM Traffic Descriptor IE are equivalent. An MPOA component must not treat the reception of a SETUP message without the Minimum Acceptable ATM Traffic Descriptor IE as an error.

4.8.9.2 QoS Parameter

Class 0 is the only class allowed when using the UBR Service Category. The use of Classes other than Class 0, is outside the scope of this specification.

Use

This IE must be included in a SETUP message.

Format

Field	Value
QoS Class Forward	0
QoS Class Backward	0

UNI 3.0 considerations

In UNI 3.0, the two-bit coding standard field is set to “00”. In UNI 3.1 and later, this field is set to “11” as the ITU-T has now standardized QoS class 0. MPOA components should treat both values as equivalent.

UNI 4.0 considerations

UNI 4.0 allows, for some ATM service categories, the signalling of individual QoS parameters for the purpose of giving the network and called party a more exact description of the desired delay and cell loss characteristics. This capability uses the End to End Transit Delay IE and the Extended QoS Parameters IE. However UNI 4.0 does not allow these IEs to be used with the UBR service category, and the only QoS Class allowed for UBR is Class 0. See Table A9-2 in [UNI4.0] for the allowed combinations of QoS parameters and ATM service categories.

4.8.9.3 AAL5 Parameters

The only AAL5 parameter that may be negotiated is the CPCS-SDU size. This is the maximum number of octets that may be transferred in an AAL5 frame on that VCC. This includes any Data-Link Layer octets (e.g. MAC and LLC sub-layers). The term “MTU size” refers to the size of the internetwork layer PDU (e.g. NHRP packet starting with ar$afn byte, or IP packet starting with the version number field). The MTU size does not include any MAC or LLC octets.

For MPOA components, the default CPCS-SDU size is 1536 octets. All MPOA components must support initiating and receiving VCCs that use a CPCS-SDU size of the default size in both the forward and backward directions.

An MPOA component may attempt to negotiate a larger CPCS-SDU size, in accordance with the procedures specified in Annex F of [UNI 3.1]. When doing so, the proposed value of CPCS-SDU must not be less than the default size. When an MPOA component proposes the use of a larger size than the default, it must be prepared to accept the remote party negotiating the value downwards. If an MPOA component receives an incoming call that proposes a larger value than the default, it must not treat this an error. Instead, it may accept the proposed value if it can support that, or use a smaller value that must not be less that the default value. The negotiated value is not restricted to being one of the LANE maximum frame sizes. For example two MPOA hosts could negotiate a size of 65535 bytes. If a remote party negotiates the CPCS-SDU down to a value lower than the default size, an MPOA component may use choose to use the VCC (performing any necessary fragmentation) or to release the VCC.

If an MPOA/NHRP Resolution exchange was used to obtain the remote ATM address to which an MPOA component wishes to set up an SVC, then the desired MTU of the remote party should be known in advance of SVC establishment, as an MPOA component is required to include a non-zero value for desired MTU size in an MPOA Cache Imposition Reply. If there is no explicit indication of MTU size in an MPOA/NHRP Resolution Reply then a value of 9180 should be assumed as the desired MTU size of the remote party. If LLC/SNAP is to be used on the SVC then the CPCS-SDU size should be at least 8 bytes greater than the MTU size, to allow space for the LLC/SNAP header. If the SVC is to be used for data packets using the MPOA tagged encapsulation the CPCS-SDU size should be at least 12 bytes greater than the MTU size, to allow space for the MPOA tag and LLC/SNAP header.

If an MPOA/ NHRP Resolution exchange was not used to obtain the remote ATM address to which an MPOA component wishes to set up an SVC, then the default CPCS-SDU size appropriate to the type of interface (LANE or Native ATM) over which the destination is reached should be used. Typically this will be the case for SVCs that are setup to adjacent Components on the same ELAN/LIS for the purpose of transferring Control PDUs. For an ELAN the MPOA default value of 1536 should be proposed, independent of the maximum frame size or emulation type in use on that ELAN. One reason for this is that many LECs, with different parameters, may be served by a single MPC. This implies that an edge device that only supports Token-Ring, for example, may be required to send its control flows over SVCs that use the MPOA default size. For Native IP over ATM the CPCS-SDU default size is 9188 (9180 MTU + 8 LLC/SNAP). Thus for example if an MPS is setting up an SVC to a next hop NHS reached over a Native ATM interface it should propose a CPCS-SDU size of 9188.

The CPCS-SDU sizes that an MPOA edge device will wish to use are determined by the maximum frame sizes of the LAN media attached to the edge device. An MPC is not required to do fragmentation of internetwork layer packets. If after an exchange of NHRP packets an MPC determines that to meet the MTU capabilities of the remote party it must perform fragmentation, it may choose to keep using the default hop-by-hop LANE path and not establish a shortcut. An intermediate router will perform any necessary fragmentation. Note that if possible all hosts in an MPOA network, both LAN and ATM attached, should use mechanisms such as Path MTU Discovery to reduce or eliminate the requirement for the network to fragment packets, as the inefficiencies due to performing fragmentation may be significant.

For MPOA, the AAL5 adaptation layer IE must always contain values for both the Forward and Backward CPCS-SDU sizes, and the IE itself must be included in both SETUP and CONNECT messages. The requirement to include the AAL5 adaptation layer IE in a CONNECT message is an additional requirement over the Annex F of [UNI 3.1] procedures, but is required by [RFC 1755].

The SSCS type parameter should be omitted. If present it must be coded to indicate the null SSCS (a value of zero). An MPOA component must treat the presence of this parameter with a zero value, and the absence of the parameter as equivalent.

Use

This IE must be present in both SETUP and CONNECT messages.

Format

Field	Value
AAL Type	5
Forward Max SDU Size ID	140
Forward Max SDU Size	as desired
Backward Max SDU Size ID	129
Backward Max SDU Size	as desired

UNI 3.0 considerations

For UNI 3.0, the mode parameter should be omitted. If included it should be set to 1 to indicate message mode. This parameter must be ignored by an MPOA component. For UNI 3.1 and later the mode parameter is illegal.

The value for the cause "AAL Parameter Cannot be Supported" is 93 in UNI 3.0 and 78 in UNI 3.1. (Note that the value of 78 for this error indication in UNI 3.1 is an error and the correct value should have been 93).

MPOA & NHRP-only interoperability considerations

If a NHRP-only Component does not support CPCS-SDU size negotiation then it will not be possible to establish an SVC between it and an MPOA component that does not support the NHRP default size. This situation may be recognized if an MPOA component initiates an SVC that gets rejected with the error code for "AAL5 parameters cannot be supported", or an MPOA component accepts an incoming call but negotiates the CPCS-SDU value down, and then finds that the call is immediately released by the remote party, with the same error code. An MPOA component should recognize this situation and stop attempting to establish a VCC to those Components that exhibit this behavior.

4.8.9.4 B-LLI

The purpose of the B-LLI IE is to provide a means to be used for compatibility checking by an addressed entity. It specifies the layer 2 and/or layer 3 protocol, and thus the encapsulation, that the calling party plans to use on the VCC being established.

LLC/SNAP encapsulation is the default encapsulation for MPOA, and this encapsulation must be implemented by all MPOA components. The use of B-LLI negotiation is allowed, but is not required. An MPOA component must include a B-LLI IE in a SETUP message encoded as shown in the table below. The layer 2 information indicates LLC and there is no layer 3 information.

An MPOA component may implement B-LLI negotiation procedures as defined in Annex C of [UNI 3.1]. In this way an encapsulation other than the default may be negotiated and used. An example of an alternative encapsulation is the null encapsulation (VCC-multiplexing), as described in [RFC 1483], where an internetwork layer packet is carried in the AAL5 CPCS-PDU payload, with no Data-Link Layer information included. Up to three instances of the B-LLI IE may be included in a SETUP message. The order of appearance of the B-LLI IEs indicates the order of preference, i.e. most favored B-LLI is first. If this negotiation is used, one of the B-LLI IEs must indicate the use of LLC, as described above. An MPOA component receiving a SETUP must not assume that the B-LLI IE indicating LLC is the first or only B-LLI IE in the SETUP message. This applies even if the MPOA component does not support the use of any encapsulation other than LLC/SNAP. This allows a station that proposes negotiation to interoperate with one that does not.

A single B-LLI IE may be included in the CONNECT message. If it is not included, it means that the called party accepts the first (or only) B-LLI IE in the SETUP message. If it is included, it means that the called party is explicitly indicating the B-LLI IE it is accepting.

Use

This IE must be used in a SETUP message. It may be used in a CONNECT message.

Format

Field	Value
Layer 2 ID	2
User Information Layer 2 Protocol	12 LAN LLC (ISO 8802/2)

4.8.9.5 Broadband Bearer Capability

An MPOA component must support the use of both BCOB-C and BCOB-X. It must be able to signal either capability, and receive an incoming call that uses either capability. If the call is rejected due a problem with the bearer capability (causes 57 - not authorized, 58 - not presently available, or 65 - not implemented), it should retry using the other value.

When a user specifies BCOB-C, the user is requesting more than an ATM-only service. The network may look at the AAL and provide service based on it. When the user specifies BCOB-X, the user is requesting an ATM-only service from the network, and the network shall not process any higher layer protocols (e.g. AAL protocols). Note that in UNI 4.0 the frame discard feature can be used with both BCOB-C and BCOB-X.

This specification recommends that BCOB-X be the default, with BCOB-C used when configured to do so, or when a call using BCOB-X failed due to the causes listed above.

Appendix F of [UNI 3.1] specifies additional rules for the encoding of this IE, that supplement those specified in section 5.4.5.7 of [UNI 3.1]. Note that octet 5a must be absent if BCOB-C is used.

Use

This IE must be present in a SETUP message.

Format

Note: the table shows the default value, not the only valid one

Field	Value
Bearer Class	16 (BCOB-X)
Traffic Type	0 (no indication)
Timing Requirements	0 (no indication)
Susceptibility to Clipping	0 (not susceptible)
User Plane Connection Configuration	0 (point to point)

4.8.9.6 ATM Addressing information

Addressing information is conveyed using the following IEs:

- Calling Party Number
- Called Party Number
- Calling Party Subaddress
- Called Party Subaddress

An MPOA component must support the use of private ATM addresses, and must support the use of all three formats of private ATM addresses (DCC format, ICD format, E.164 format). An MPOA component may support the use of Native E.164 ATM addresses (for which there is a single format). The control ATM address of an MPOA component must be a private ATM address.

An MPOA component is not required to support the use of the Calling Party Subaddress or Called Party Subaddress IEs. These IEs are used to carry a private ATM address across a public network that supports only Native E.164 Addresses. LANE does not require the use of the Calling Party Subaddress or Called Party Subaddress IEs and MPOA makes no additional requirements in this regard. An MPOA component may support the use of the Calling Party Subaddress or Called Party Subaddress IEs.

Use

The Calling Party Number and Called Party Number IEs must be used in a SETUP message. The Calling Party Subaddress and Called Party Subaddress IEs may be used in a SETUP message.

Format

The formats for private ATM addresses are shown. The coding of Calling/Called Party Number using Native E.164 ATM addresses and the coding of the Subaddress IEs are as defined in [UNI 3.1].

Calling Party Number

Field	Value
Type of Number	0 (unknown)
Addressing/Numbering Plan Identification	2 (ATM endsystem address)
Presentation Indicator	0 (presentation allowed)
Screening Indicator	1 (user provided verified & passed)
Address Octets	address value

Called Party Number

Field	Value
Type of Number	0 (unknown)
Addressing/Numbering Plan Identification	2 (ATM endsystem address)
Address Octets	address value

UNI 3.0 considerations

In UNI 3.1, the ATM Endsystem Address type was introduced to differentiate ATM addresses from OSI NSAPs. In UNI 3.0, ‘ATM Endsystem Address’ is not a valid type. Therefore, in the Called and Calling Party Subaddress IEs, the three-bit ‘type of subaddress’ field must specify ‘NSAP’ (value = 001) when using the Subaddress IE to carry ATM addresses.

5. MPOA Frame Formats [Normative]

5.1 Encapsulation

5.1.1 Data Frame Encapsulation

By default, MPOA uses LLC encapsulation for all data flows in accordance with the rules defined in [RFC 1483]. The default shortcut data encapsulation is the RFC 1483 LLC Encapsulation for Routed Protocols, shown in Figure 11. MPOA allows the negotiation of alternative encapsulations (e.g. Null Encapsulation) using the B-LLI negotiation procedures defined in ATM Signalling [UNI 3.0,UNI 3.1, UNI 4.0]. All MPOA devices must be able to use the default encapsulation for all data flows on all VCCs. Negotiation of other encapsulations is optional.

MPOA also allows the optional use of the MPOA Tagged Encapsulation shown in Figure 12 for data flows. The format for tagged packets consists of an 8-byte LLC/SNAP header, a 4-byte tag field, and the Internetwork layer packet. The tag field must be used to determine the Internetwork layer protocol type of the packet which follows. There may be a mix of tagged and non-tagged packets on a VCC.

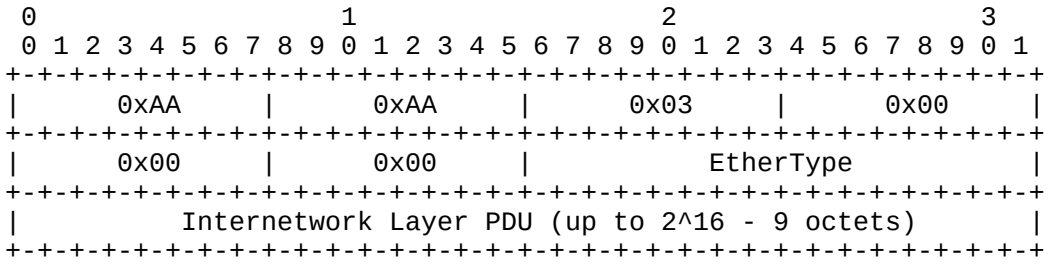


Figure 11 RFC 1483 LLC Encapsulation for Routed Protocols

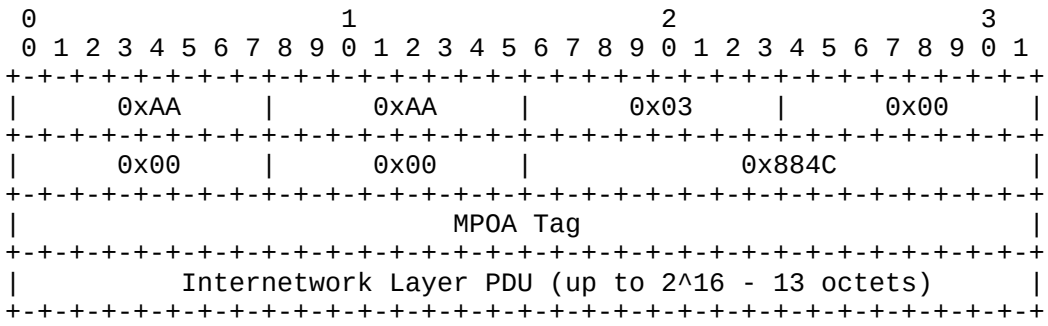


Figure 12 MPOA Tagged Encapsulation for Routed Protocols

5.1.2 Control Frame Encapsulation

By default, MPOA uses LLC encapsulation for all control flows as defined in [NHRP] and shown in Figure 13. MPOA allows the negotiation of alternative encapsulations (e.g. Null Encapsulation) using the B-LLI negotiation procedures defined in ATM Signalling [UNI 3.0,UNI 3.1, UNI 4.0].

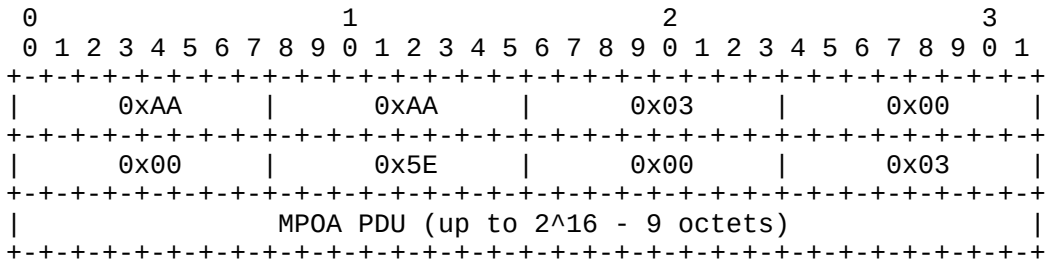


Figure 13 MPOA Control Frame Encapsulation

5.2 LANE TLVs

5.2.1 MPS Configuration TLVs

The following TLVs may be used to change the default values for MPS parameters.

TLV Name	Type	Length	Value
Keep-Alive Time	00-A0-3E-1D	2	MPS-p1in seconds
Keep-Alive Lifetime	00-A0-3E-1E	2	MPS-p2 in seconds
Internetwork-layer Protocols	00-A0-3E-1F	8	MPS-p3 encoding: Control (1 octet): 0x00=DISABLE MPOA Resolution support for the protocol. 0x01=ENABLE MPOA Resolution support for the protocol. Short protocol (2 octets): (Encoded as specified in [NHRP]). Long protocol (5 octets): (Encoded as specified in [NHRP]). *Note: Multiple Internetwork-layer Protocol TLVs may be present.*
MPS Initial Retry Time	00-A0-3E-20	2	MPS-p4, in seconds.
MPS Retry Time Maximum	00-A0-3E-21	2	MPS-p5, in seconds.
MPS Give-up Time	00-A0-3E-22	2	MPS-p6 in seconds
Default Holding Time	00-A0-3E-23	2	MPS-p7, in seconds.

5.2.2 MPC Configuration TLVs

The following TLVs may be used to change the default values for MPC Parameters.

TLV Name	Type	Length	Value
SC-Setup Frame Count	00-A0-3E-24	2	MPC-p1
SC-Setup Frame Time	00-A0-3E-25	2	MPC-p2 in seconds

Flow-detection Protocols	00-A0-3E-26	8	MPC-p3 encoding: Control (1 octet): 0x00=DISABLE flow detection for the protocol 0x01=ENABLE flow detection for the protocol. Short protocol (2 octets): (Encoded as specified in [NHRP]). Long protocol (5 octets): (Encoded as specified in [NHRP]). *Note: Multiple Flow-detection Protocols TLVs may be present.*
MPC Initial Retry Time	00-A0-3E-27	2	MPC-p4, in seconds.
MPC Retry Time Maximum	00-A0-3E-28	2	MPC-p5, in seconds.
Hold Down Time	00-A0-3E-29	2	MPC-p6, in seconds.

5.2.3 Device Type TLV

The MPOA Device Type TLV contains two sub-fields within the value field, as shown below. One sub-field identifies the type of MPOA device (server or client). The other sub-field specifies the ATM address of the MPOA device using UNI 4.0 encoding.

TLV Name	Type	Length	Value
MPOA Device Type	00-A0-3E-2A	N	MPOA DEVICE TYPE (1 octet) 0 == Non-MPOA device 1 == MPOA Server 2 == MPOA Client 3 == MPS and MPC 4-255 undefined (reserved for future use) Number of MPS MAC Addresses (1 octet) This field is always present. Usage depends on device type, as follows: device type | use 0 | must be zero (i) 1 | may be zero or non-zero (ii) 2 | must be zero (iii) 3 | must be non-zero (iv) MPS Control ATM ADDRESS (20 octets) Present only for device types 1 and 3. (Private ATM address format. See UNI 4.0 sec. 3.0 for encoding)

			MPC Control ATM ADDRESS (20 octets) Present only for device types 2 and 3. (Private ATM address format. See UNI 4.0 sec. 3.0 for encoding) MPS MAC Addresses (Variable length)

(i) all MAC addresses served by the ATM address are non-MPOA MAC addresses.

(ii) if zero, then all MAC addresses served by the ATM address are MPS MAC addresses; if non-zero all MAC addresses served by the ATM address are non-MPOA MAC addresses except for those enumerated in the MAC address list, which are MPS MAC addresses.

(iii) all MAC addresses served by the ATM address are MPC MAC addresses.

(iv) all MAC addresses served by the ATM address are MPC MAC addresses except for those enumerated in the MAC address list, which are MPS MAC addresses.

5.3 Frame Formats

The encapsulation used for MPOA control messages is the same as is defined for NHRP control messages in [NHRP].

MPOA specifies the following control messages:

1. MPOA Resolution Request
2. MPOA Resolution Reply
3. MPOA Cache Imposition Request
4. MPOA Cache Imposition Reply
5. MPOA Egress Cache Purge Request
6. MPOA Egress Cache Purge Reply
7. MPOA Keep-Alive
8. MPOA Trigger

These messages reuse the NHRP packet formats. The NHRP packet type values 0x80 - 0x100 are administered by IANA. MPOA control messages use the values 0x80-0x87.

MPOA also uses the following NHRP control messages between MPCs and MPSs:

1. NHRP Purge Request
2. NHRP Purge Reply

5.3.1 MPOA CIE Codes

The following codes are defined for use in MPOA messages (in addition to those defined in NHRP). These codes may be used as deemed appropriate by MPOA components.

Table 5 MPOA CIE Codes

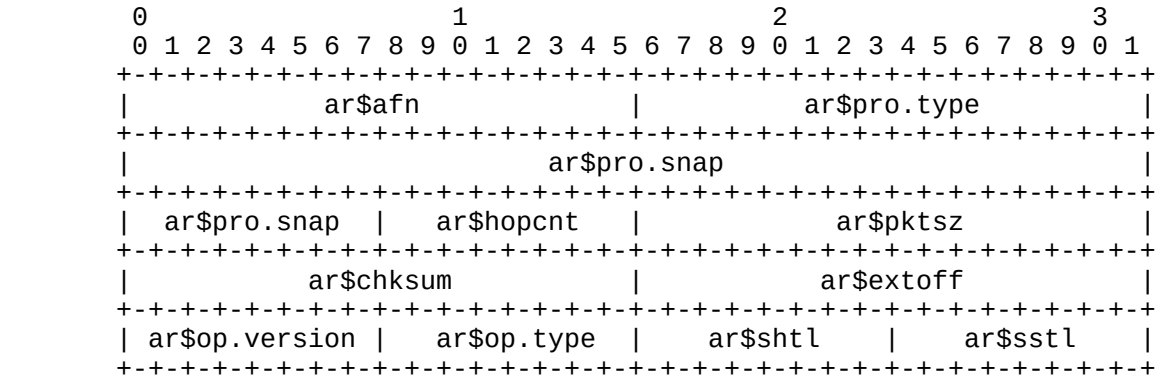
Code	Meaning
0x00	Success
0x81	Insufficient resources to accept egress cache entry.
0x82	Insufficient resources to accept shortcut
0x83	Insufficient resources to accept either shortcut or egress cache entry.
0x84	Unsupported Internetwork Layer protocol
0x85	Unsupported MAC layer encapsulation
0x86	Not an MPC
0x87	Not an MPS
0x88	Unspecified/other

5.3.2 Control Message Format

Each MPOA control message is conveyed using the NHRP packet format. NHRP Extensions may be included in each MPOA control message to convey additional information.

5.3.2.1 Fixed Header

Each MPOA control message has the same Fixed Header as an NHRP packet.



ar$op.version is set to 0x01 (NHRP).

Packet type values (ar$op.type) are assigned for MPOA control messages as follows.

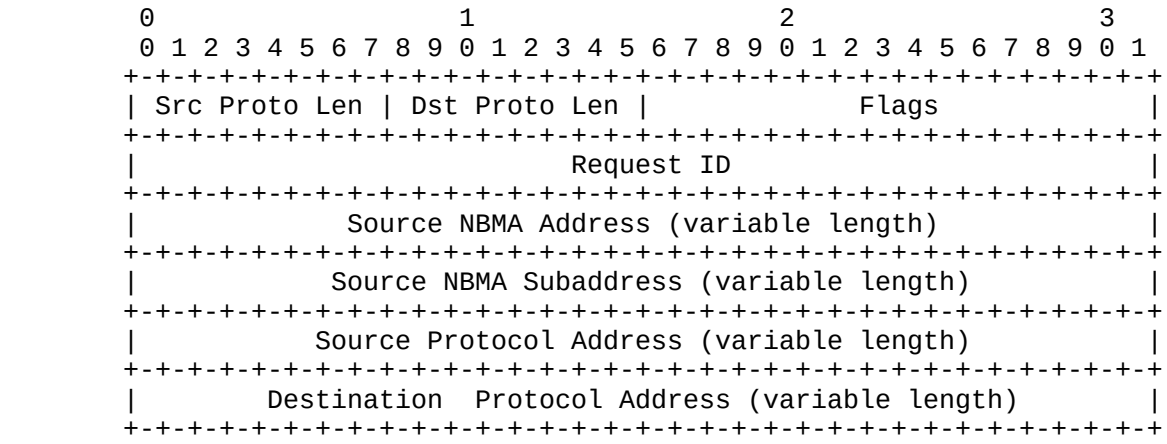
ar$op.type	**MPOA Control Message**
0x80	MPOA Cache Imposition Request
0x81	MPOA Cache Imposition Reply
0x82	MPOA Egress Cache Purge Request
0x83	MPOA Egress Cache Purge Reply
0x84	MPOA Keep-Alive
0x85	MPOA Trigger
0x86	MPOA Resolution Request
0x87	MPOA Resolution Reply

Other fields in the Fixed Header must be set in conformance with NHRP.

5.3.2.2 Common Header

Each MPOA control message has the same Common Header as an NHRP packet.

The Common Header is as follows:



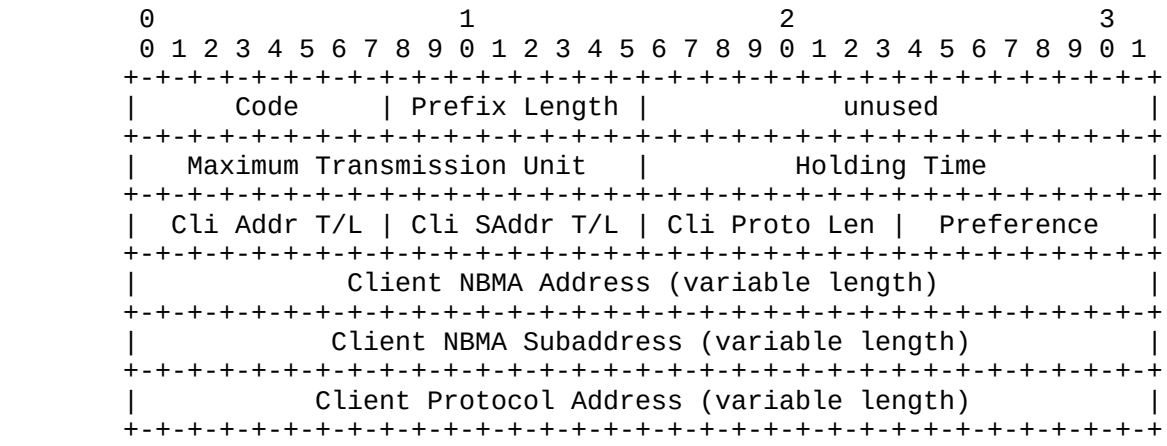
The Common Header specifies the sender's NBMA (ATM) and internetwork layer address and the receiver's internetwork layer address. Some exceptions exist in the context of certain MPOA control messages as described in the section for each message.

For consistency, all fields of the Common Header should be filled in as specified by NHRP, even though in some cases the receiving station may ignore some of them.

5.3.2.3 Client Information Element

MPOA control messages may have the same Client Information Elements as an NHRP packet.

The CIE has the following format:



The usage of the CIE for each message-specific part is described in the section for each message.

5.3.2.4 Extensions

MPOA control messages may have the same Extensions as an NHRP packet, such as Route Record, NHRP Authentication and Vendor Private Extensions.

All MPOA Extensions, summarized in the following table, use the NHRP Extension format.

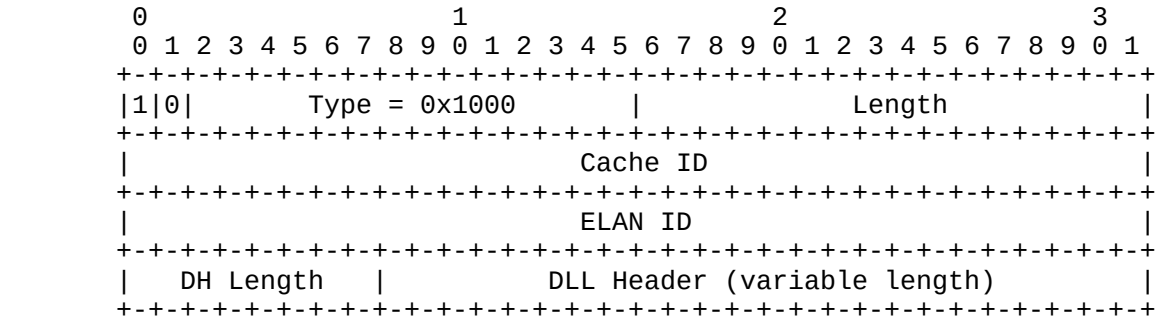
Table 6 MPOA Extensions

Type	MPOA Extension
0x1000	MPOA DLL Header Extension
0x1001	MPOA Egress Cache Tag Extension
0x1002	MPOA ATM Service Category Extension
0x1003	MPOA Keep-Alive Lifetime Extension
0x1004	MPOA Hop Count Extension
0x1005	MPOA Original Error Code Extension

5.3.2.4.1 MPOA DLL Header Extension

The MPOA DLL Header Extension is used to convey Data-Link Layer Header information [including MAC destination address, MAC Source Address, Route Information Field (in the case of 802.5 Token Ring), Ethernet Type (in the case of Ethernet), or LLC Header (in the case of IEEE 802)].

The MPOA DLL Header Extension format is as follows:



Cache ID specifies the egress MPS's ID for an egress cache entry. A value of zero for the cache ID is not allowed.

ELAN ID specifies a LANE ELAN ID.

DH Length specifies the length of the DLL header.

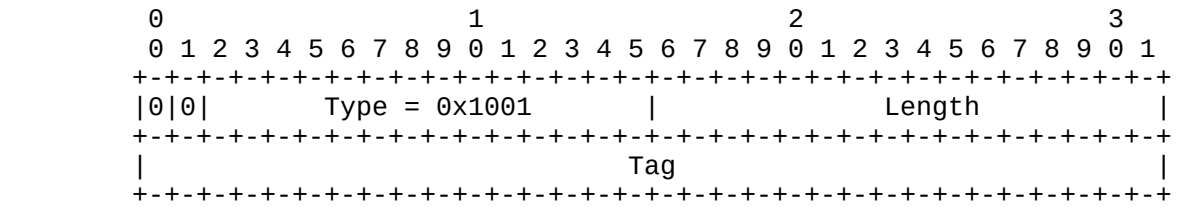
DLL header is used to specify Data-Link Layer Header information.

The specific usage of this Extension is described in the appropriate section for each message.

5.3.2.4.2 MPOA Egress Cache Tag Extension

The MPOA Egress Cache Tag Extension is used to convey egress cache tags.

The MPOA Egress Cache Tag Extension format is as follows:



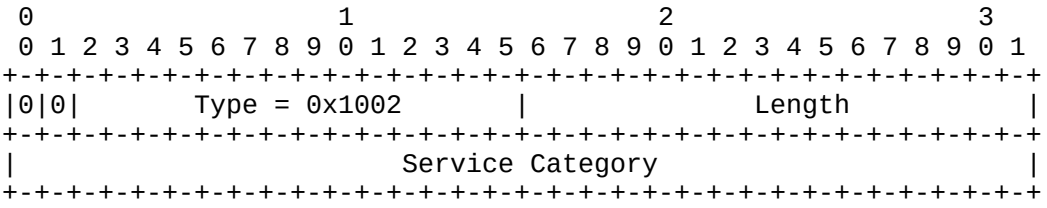
The tag is a 32 bit value chosen by the egress MPC.. A value of zero for the tag is not allowed

The specific usage of this Extension is described in the appropriate section for each message.

5.3.2.4.3 MPOA ATM Service Category Extension

The MPOA ATM Service Category Extension is used to convey the set of Service Categories supported by the sender.

The MPOA ATM Service Category Extension format is as follows:



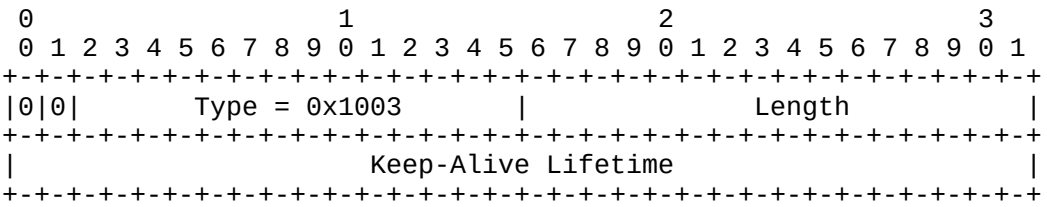
Service Category has precisely the same syntax and semantics defined in the corresponding LANE TLV [LANE].

The specific usage of this Extension is described in the appropriate section for each message.

5.3.2.4.4 MPOA Keep-Alive Lifetime Extension

The MPOA Keep-Alive Lifetime is used to convey the duration of time that a Keep-Alive message may be considered valid.

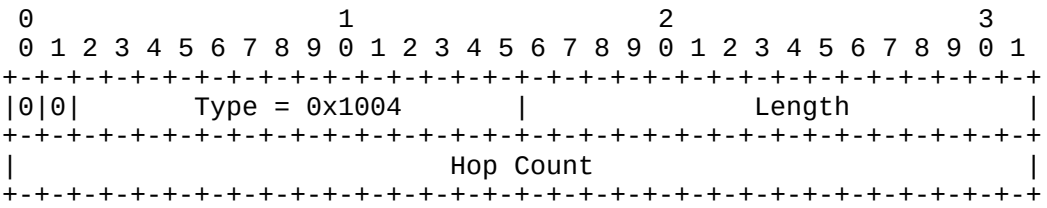
The MPOA Keep-Alive Lifetime Extension format is as follows:



5.3.2.4.5 MPOA Hop Count Extension

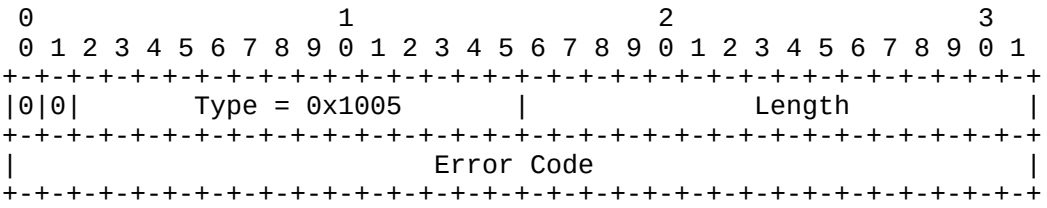
This extension is used to convey the hop count limit for forwarding internetworking packets. It is used for internetworking protocols that use a hop count field that counts up (e.g. the transport control field in IPX) and not down (e.g. the TTL field in IP). In order to be forwarded over a shortcut VCC the internetworking layer packet must contain a hop count that is less than the value specified in the hop count extension.

The MPOA Hop Count Extension format is as follows:



5.3.2.4.6 MPOA Original Error Code Extension

This extension may be included in any MPOA message. Its purpose is to preserve the original MPOA error code through conversions to and from NHRP messages. It may only be included in a Reply if the extension was included in a Request. An ingress MPC may include this extension with a null value in an MPOA Resolution Request, to allow its use in an MPOA Cache Imposition Reply. It may also be included in an MPOA Egress Cache Purge Request message so that it may be included with the subsequent NHRP Purges that the egress MPS may generate.



5.3.3 MPOA Resolution Request Format

An MPOA Resolution Request is sent from an ingress MPC to an ingress MPS to request the egress ATM address corresponding to an internetwork layer destination address. Upon receipt of an MPOA Resolution Request, an ingress MPS must send a new NHRP Resolution Request towards the egress MPS.

Fixed Header

ar$op.type is set to 0x86 (MPOA Resolution Request).

Common Header

The Common Header is coded as follows:

Field	Usage
Flags	Unused.
Source NBMA Address	The ATM address of the ingress MPC from which internetwork layer datagrams will be sent.
Source NBMA Subaddress	The ATM Subaddress of the ingress MPC from which internetwork layer datagrams will be sent.
Source Protocol Address	Optional. Source Protocol address of the MPC if used. If not used, Src Proto Len field must be set to zero and no storage is allocated for the Source Protocol address..
Destination Protocol Address	The internetwork layer address of the final destination to which the internetwork layer datagrams will be sent.

Other fields in the Common Header must be set in conformance with NHRP.

Client Information Element

One CIE may be added in conformance with NHRP.

Field	Usage
Prefix Length	Largest Acceptable prefix length.
MTU	Unused and must be set to zero (0).

Extensions

An MPOA Egress Cache Tag Extension must be added as follows:

Field	Usage
Type	This field must be set to 0x1001 for an MPOA Egress Cache Tag Extension.
Length	The length is set to zero (0) and no storage is allocated for the tag.

An MPOA ATM Service Category Extension should be added as follows:

Field	Usage
Type	This field must be set to 0x1002 for an MPOA ATM Service Category Extension.
Length	This field is coded as specified by NHRP.

Service Category	The Service Category is set to indicate the Service Categories supported by the ingress MPC.

5.3.4 MPOA Resolution Reply Format

An MPOA Resolution Reply is sent from an ingress MPS to an ingress MPC in reply to a corresponding MPOA Resolution Request upon receiving an NHRP Resolution Reply from the egress MPS.

Fixed Header

ar$op.type is set to 0x87 (MPOA Resolution Reply).

Common Header

The Common Header is coded as follows:

Field	Usage
Flags	Unused.
Source NBMA Address	The ATM address of the ingress MPC from which internetwork layer datagrams will be sent.
Source NBMA Subaddress	The ATM Subaddress of the ingress MPC from which internetwork layer datagrams will be sent.
Source Protocol Address	Copied from corresponding MPOA Resolution Request.
Destination Protocol Address	The internetwork layer address of the final destination to which the internetwork layer datagrams will be sent.

Other fields in the Common Header must be set in conformance with NHRP.

Client Information Element

A CIE must always be present in the MPOA resolution reply.

For a normal resolution reply, all CIEs must be copied from the corresponding NHRP Resolution Reply.

To report an error condition the ingress MPS must include a CIE as follows:

Field	Usage
Code	Selected CIE Code
Prefix Length	Unused.
Maximum Transmission Unit	Unused.
Holding Time	Unused.
Cli Addr T/L	This field must be set to zero and no storage is allocated for the Client NBMA Address.
Cli SAddr T/L	This field must be set to zero and no storage is allocated for the Client NBMA Subaddress.
Cli Proto Len	This field must be set to zero and no storage is allocated for the Client Protocol Address.

Extensions

All Extensions must be copied from the corresponding NHRP Resolution Reply, except for those extensions the ingress MPS added to the initial request.

5.3.5 MPOA Cache Imposition Request Format

An MPOA Cache Imposition Request is sent from an egress MPS to an egress MPC to impose an egress cache entry upon receipt of an NHRP Resolution Request from the ingress MPS.

Fixed_Header
ar$op.type is set to 0x80 (MPOA Cache Imposition Request).

Common_Header
The Common Header is coded as follows:

Field	Usage
Flags	Unused
Request ID	This field is a Request ID for the imposition transaction. The MPS should assign a Request ID for the MPOA Cache Imposition Request that is unique over all unacknowledged impositions.
Source NBMA Address	This field must be copied from the NHRP Resolution Request. This is the ATM address of the ingress MPC (or NHC) from which internetwork layer datagrams will be sent.
Source NBMA Subaddress	This field must be copied from the NHRP Resolution Request. This is the ATM Subaddress of the ingress MPC (or NHC) from which internetwork layer datagrams will be sent.
Source Protocol Address	This field is set to the internetwork layer address of the egress MPS.
Destination Protocol Address	This field must be copied from the NHRP Resolution Request. This address refers to the internetwork layer address of the final destination to which the internetwork layer datagrams will be sent.

Other fields in the Common Header must be set in conformance with NHRP.

Client_Information_Element
The CIE is coded as follows:

Field	Usage
Code	Unused.
Prefix Length	Prefix length in bits for the Destination Protocol Address, as known to the egress MPS.
Maximum Transmission Unit	The egress MPS must set this field based on either local information, or copy it from the NHRP Resolution Request.
Holding Time	The number of seconds for which this entry should be considered to be valid in the egress cache. The value given here must be at least twice as large as the one that will be returned in the corresponding NHRP Resolution Reply.
Cli Addr T/L	This field must be set to zero and no storage is allocated for the Client NBMA Address.
Cli Saddr T/L	This field must be set to zero and no storage is allocated for the Client NBMA Subaddress.
Cli Proto Len	This field must be set to zero and no storage is allocated for the Client Protocol Address at all.
Preference	Unused.

Extensions
All NHRP Extensions must be copied from the original NHRP Resolution Request.

If the holding time in the CIE is non-zero, an MPOA DLL Header Extension must be included as follows:

Field	Usage
Type	This field must be set to 0x1000 for an MPOA DLL Header Extension.
Length	This field is coded as specified by NHRP.

Cache ID	This field is used to convey cache ID to be set in the egress cache. Cache ID value must be selected by the egress MPS so that a different ID is assigned for each egress cache entry.
ELAN ID	This field is set to the LANE ELAN Id.
DH Length	The length of the DLL Header.
DLL Header	The DLL header to be used for encapsulating internetwork layer datagrams when the egress MPC receives the datagrams from the shortcut before sending them to the higher layers.

The MPS sends the egress MPC the DLL header that the egress router would use to transmit frames along the default routed path via LANE to the destination specified in the NHRP common header.

For a given LAN type, the DLL headers are self-describing. It is the responsibility of the egress MPC to parse the DLL header provided by the MPS to determine whether the given encapsulation is supported for the given protocol. If the MPC does not support the encapsulation or protocol provided in the MPOA Cache Imposition Request, the MPC must return a status value in the MPOA Cache Imposition Reply indicating that either the DLL encapsulation or protocol is not supported.

For DLL encapsulations that contain a length field, such as 802.3 LLC SNAP, the length field in the DLL header section of the MPOA Cache Imposition Request must be filled in with the correct length for an empty frame. In the 802.3 LLC SNAP case, for example, the length field is set to 8.

The MPOA Imposition Request message is also used to purge egress cache entries in the egress MPC. In this case the source NBMA Address field must be NULL, the holding time field must be set to zero and MPOA DLL header extension is optional.

5.3.6 MPOA Cache Imposition Reply Format

An MPOA Cache Imposition Reply is sent from an egress MPC to an egress MPS in reply to an MPOA Cache Imposition Request.

Fixed Header

ar$op.type is set to 0x81 (MPOA Cache Imposition Reply).

Common Header

The Common Header must be copied from the corresponding MPOA Cache Imposition Request.

Client Information Element

CIEs are coded as follows:

Field	Usage
Code	Selected CIE Code
Prefix Length	Actual prefix length imposed in bits for the Client Protocol Address. The prefix length may be set to indicate a host route.
Maximum Transmission Unit	The MTU size should be set to the maximum value allowed by the egress MPC. This value must be non-zero.
Holding Time	Unused.
Cli Addr T/L	This field must be set to zero (and no Client ATM Address included) unless the status is Success.
Cli Saddr T/L	This field must be set to zero (and no Client ATM Subaddress included) unless the status is Success.
Cli Proto Len	This field must be set to zero and no storage is allocated for the Client Protocol Address at all.

Client NBMA Address	ATM address of the egress MPC that will receive internetwork layer datagrams via the shortcut.
Client NBMA Subaddress	ATM subaddress of the egress MPC that will receive internetwork layer datagrams via the shortcut (if any).

Extensions

All Extensions must be copied from the corresponding MPOA Cache Imposition Request.

If an MPOA Egress Cache Tag Extension is included, it must be set as follows:

Field	Usage
Length	If a valid tag exists, the length is set to four (4).
Tag	Set to the value of the tag chosen by the egress MPC.

If an MPOA Service Category Extension is included, it must be set as follows:

Field	Usage
Service Category	The Service Category is set to indicate the Service Categories supported by the egress MPC.

5.3.7 MPOA Egress Cache Purge Request Format

An MPOA Egress Cache Purge Request is sent from an egress MPC to an egress MPS to purge an egress cache entry. .

Fixed Header

ar$op.type is 0x82 (MPOA Egress Cache Purge Request).

Common Header

The Common Header is coded as follows:

Field	Usage
Flags	The Flags field is coded as follows: 0 1 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+- |N| unused | +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+- N: No-Reply Flag
Request ID	The semantics of this field are the same as that of NHRP. This value must be selected by the egress MPC, so that it may recognize the corresponding MPOA Egress Cache Purge Reply.
Source NBMA Address	Data ATM Address of the egress MPC.
Source NBMA Subaddress	Data ATM Subaddress of the egress MPC (if any).
Source Protocol Address	Internetwork layer address of the MPC (if any). If the MPC has no internetwork layer address, Src Proto Len must be set to zero and no storage is allocated for the Source Protocol Address.
Destination Protocol Address	Internetwork layer address of the egress MPS.

Other fields in the Common Header must be set in conformance with NHRP.

Client Information Element

CIEs are coded as follows:

Field	Usage
Prefix Length	Destination Prefix Length.
Client Protocol Address	Destination Protocol Address (to purge)
Client NBMA Address	Egress MPC Data ATM Address
Client NBMA Subaddress	Egress MPC Data subaddress (if any)

Extensions
An MPOA DLL Header Extension must be included as follows:

Field	Usage
Cache ID	This field is used to convey a cache ID of the egress cache entry being invalidated.
ELAN ID	Unused.
DH Length	This field must be set to zero and no storage is allocated for the DLL Header.

5.3.8 MPOA Egress Cache Purge Reply Format

An MPOA Egress Cache Purge Reply is sent from an egress MPS to an egress MPC in reply to an MPOA Egress Cache Purge Request.

An MPOA Egress Cache Purge Reply is formed from an MPOA Egress Cache Purge Request by changing the ar$op.type to 0x83.

Fixed Header
ar$op.type is set to 0x83 (MPOA Egress Cache Purge Reply).

Common Header
The Common Header must be copied from the corresponding MPOA Egress Cache Purge Request.

Client Information Element
All CIEs must be copied from the corresponding MPOA Egress Cache Purge Request.

Extensions
All Extensions must be copied from the corresponding MPOA Egress Cache Purge Request.

5.3.9 MPOA Keep-Alive Format

An MPOA Keep-Alive is periodically sent from an MPS to an MPC(s).

Fixed Header
ar$op.type is set to 0x84 (MPOA Keep-Alive).

Common Header
The Common Header is coded as follows:

Field	Usage
Flags	Unused.
Request ID	The sequence number of the Keep-Alive Message. This value must be set to zero on the first transmission and must be incremented by at least one each time the MPS sends this message to a given MPC.
Source NBMA Address	Control ATM Address of the MPS.
Source NBMA Subaddress	NULL
Source Protocol Address	NULL

Destination Protocol Address	NULL

Client Information Element
There are no CIEs for this message.

Extensions
An MPOA Keep-Alive Lifetime Extension must be added as follows:

Field	Usage
Type	This field must be set to 0x1003 for an MPOA Keep-Alive Lifetime Extension.
Length	Two (2) Octets
Keep-Alive Lifetime	Set to the duration of time that a Keep-Alive message may be considered valid

5.3.10 MPOA Trigger Format

An MPOA Trigger is sent from an ingress MPS to an ingress MPC to request the ingress MPC to issue MPOA Resolution Requests.

Fixed Header
ar$op.type is set to 0x85 (MPOA Trigger).

Common Header
The Common Header is coded as follows:

Field	Usage
Flags	Unused.
Request ID	Unused (there is no reply/ACK packet for this message).
Source NBMA Address	Control ATM address of the ingress MPS.
Source NBMA Subaddress	NULL
Source Protocol Address	Internetwork layer address of the ingress MPS. This must be the internetwork layer address that would be returned in an NHRP Responder Address Extension or NHRP Forward/Reverse Transit Record Extension included by that MPS (if such Extensions are included) in an NHRP Resolution Reply.
Destination Protocol Address	Destination Protocol Address (to trigger)

Client Information Element
No CIE is used.

Extensions
No extensions are used.

5.3.11 NHRP Purge When Used on the Data Plane

An NHRP Purge message may be sent on the data plane by an egress MPC to an ingress MPC or NHC to purge ingress cache entries.

Fixed Header
ar$op.type is set to 5 (NHRP Purge Request).

Common Header
The Common Header is coded as follows:

Field	Usage

Flags	The Flags field is coded as follows: 0 1 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+ |N| unused | +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+ N: No-Reply Flag The N bit must be set to one.
Request ID	Unused. Must be set to zero (0).
Source NBMA Address	Egress MPC data ATM address.
Source NBMA Subaddress	Egress MPC data ATM subaddress (if any).
Source Protocol Address	1. Set to the protocol address of the egress MPS (if known) or NULL (Src Proto Len=0 and no storage is allocated for the Source Protocol Address).
Destination Protocol Address	NULL

Other fields in the Common Header must be set in conformance with NHRP.

Client Information Element

The purge request includes one or more CIEs as follows:

Field	**Usage**
Code	Set to 0x00.
Prefix Length	Prefix Length in bits for the Client Protocol Address(es) being purged. Only ingress cache entries associated with the shortcut over which the MPOA Data Plane Purge is received are purged. All ingress cache entries for this shortcut must be purged if the Prefix Length is set to 0x00.
Maximum Transmission Unit	Unused.
Holding Time	Unused.
Cli Addr T/L	This field must be set to zero and no storage is allocated for the Client NBMA address at all.
Cli SAddr T/L	This field must be set to zero and no storage is allocated for the Client NBMA Subaddress at all.
Cli Proto Len	The length in octets of the Client Protocol Address.
Client NBMA Address	Egress MPC data ATM address.
Client Protocol Address	The internetwork layer address that is being purged from the ingress MPC's cache.

Extensions

NHRP Authentication and Vendor Private Extensions may be added as desired.

6. References

[AIW] APPN Implementers' Workshop, Document ATM-09, HPR Extensions for ATM Networks.

[CLIP] M. Laubach, RFC 1577, Classical IP and ARP over ATM, January, 1994.

[IEEE 802.1d] ISO/IEC 10038; ANSI/IEEE Std. 802.1d, "Information Processing Systems - Local Area Networks - (MAC Bridges).

[LANE] Keene, J.. (editor), "LAN Emulation Over ATM Version 2 - LUNI Specification," ATM Forum (AF-LANE-0084.000), 1997.

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

[MTU] R. Atkinson, RFC 1626, Default IP MTU for use over ATM AAL5, May 1994

[NHRP] See Annex C.

[PATH MTU] J. Mogul and S. Deering, RFC 1191, Path MTU Discovery, November, 1990.

[RFC 791] J. Postel, RFC 791, Internet Protocol, September 1, 1981.

[RFC 1483] J. Heinanen, RFC 1483, Multiprotocol encapsulation over ATM Adaptation Layer 5.

[RFC 1755] M. Perez, F. Liaw, A. Mankin, E. Hoffman, D. Grossman, A. Malis, RFC 1755 ATM Signaling support for IP over ATM.

[ROUTER REQ] RFC 1812, Fred Baker, Requirements for IP Version 4 Routers, June 22, 1995.

[RSVP] B. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, Internet Draft, draft-ietf-rsvp-spec-14.txt.

[UNI 3.0] "ATM User-Network Interface (UNI) Specification, Version 3.0", The ATM Forum Technical Committee.

[UNI 3.1] "User-Network Interface (UNI) Specification, Version 3.1", The ATM Forum Technical Committee.

[UNI 4.0] "ATM User-Network Interface (UNI) Signaling Specification, Version 4.0", The ATM Forum Technical Committee.

Annex A. Protocol-Specific Considerations [Normative]

Each internetwork layer protocol provides addressing and forwarding functions in a different way, and uses different encapsulations and different demultiplexing points (e.g. LSAPs, OUIs, PIDs, and Ethernet Types) on LANs. Many internetwork layer protocols even have multiple encapsulations on a single LAN. Because of these differences, MPOA components require internetwork layer protocol-specific knowledge to perform flow detection, address resolution, and shortcut data transformations. Flow detection and address resolution require at least a minimal understanding of internetwork layer addresses. Ingress MPC and egress MPC shortcut transformations must be defined on a protocol-specific basis. Each of these functions must be specified individually for each internetwork layer protocol supported.

A.1 IP Packet Handling in MPOA

This section describes the processing of IP packets in MPOA. Note that MPOA does not define the handling of IP Packets that are not sent over shortcuts.

A.1.1 Requirements

The MPOA System must support the IP version 4 Router Requirements [ROUTER REQ]. MPOA distributes this responsibility across MPOA components.

A.1.2 Encapsulation

There are three standard formats for IP packets carried over Ethernet and Token Ring (shown in Figure 14-Figure 16), and three standard formats for IP packets carried over an MPOA shortcut (shown in Figure 17-Figure 19).

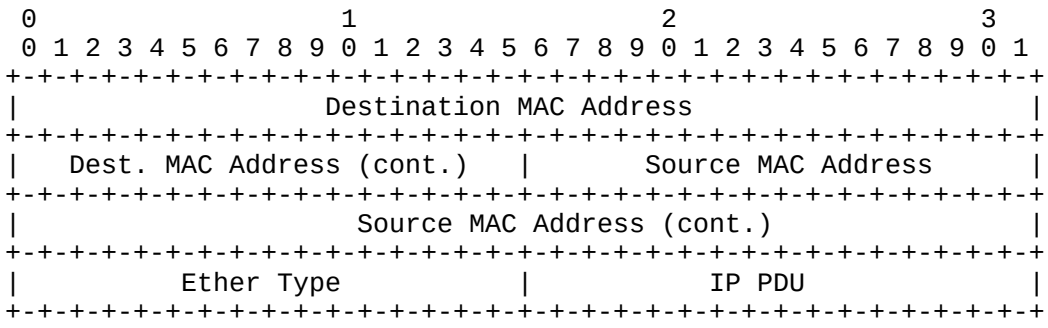


Figure 14 Ethernet IP Encapsulation

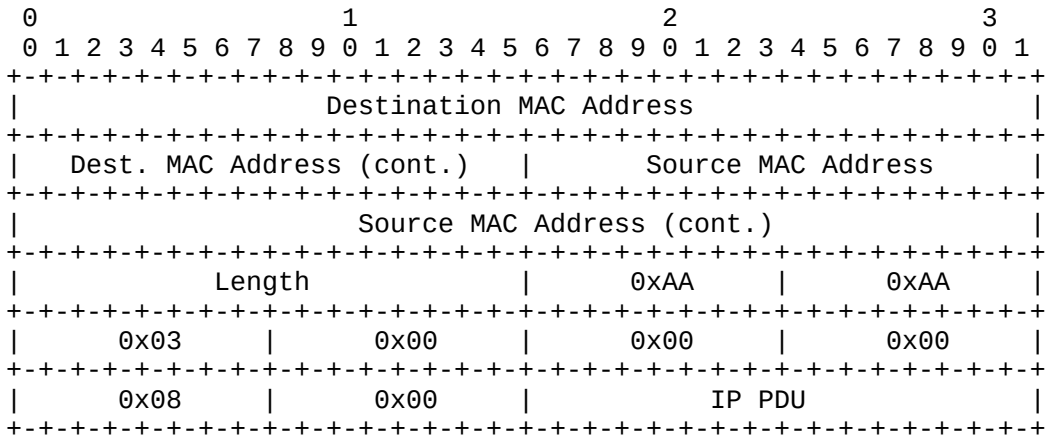


Figure 15 802.3 IP Encapsulation

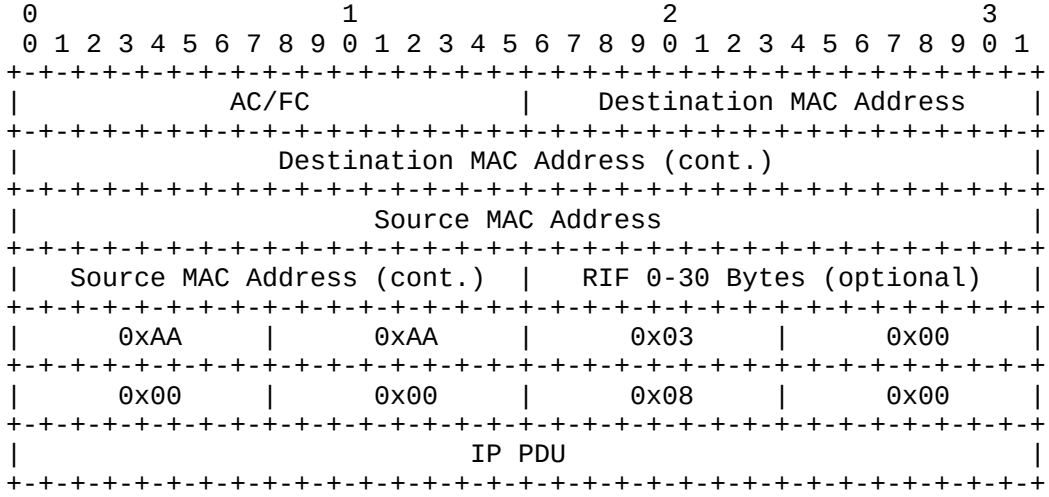


Figure 16 802.5 IP Encapsulation

The DLL header supplied in the egress cache entry consists of the entire media specific encapsulation through Ethernet Type. Note that the LECID is not included in this DLL header. Note that the 802.3 DLL header contains a length field that must be updated by the egress MPC for each packet received on a shortcut.

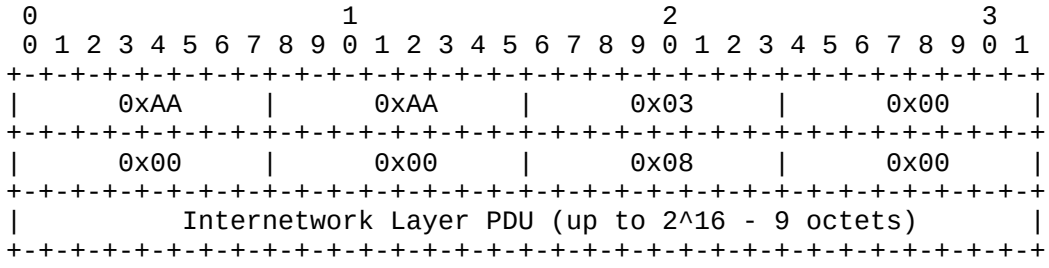


Figure 17 RFC 1483 LLC/SNAP Encapsulation for Routed IP PDUs

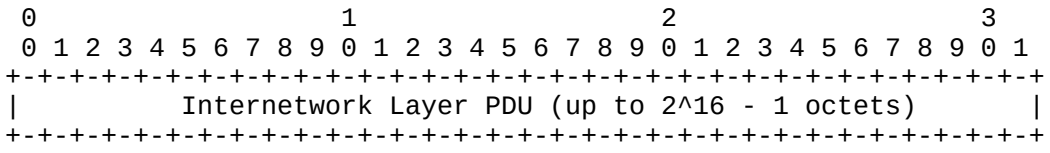


Figure 18 RFC 1483 "Null" Encapsulation for Routed IP PDUs

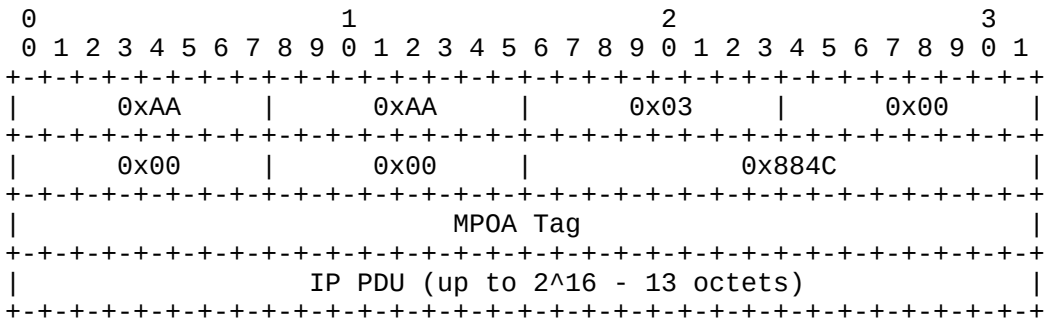


Figure 19 MPOA Tagged Encapsulation for IP

A.1.3 MPS Role

An MPS is co-located with a router and must process IP packets in conformance with [ROUTER REQ].

A.1.4 Ingress MPC Role

An MPC must perform basic IP forwarding. Additional router features may be implemented in an MPC, as described in the following subsections.

IP Options

There are some cases when the ingress MPC does not have all of the information that is required to meet the router requirements. For example, when the ingress MPC receives a Strict Source Route option in an IP packet, it does not have access to the routing table to determine whether the next hop given in the option is indeed the current next hop to the destination. Thus, the ingress MPC cannot determine on its own whether the packet is to be forwarded or dropped.

If the ingress MPC cannot correctly process an IP option, it must send the packet unmodified to the MPS via LANE. An ingress MPC may send all packets with IP options to the MPS. Of course, if an MPC has full support for an IP option, it should process the option on its own and send the packet over a shortcut.

TTL

When sending a packet via LANE, the MPC must not modify the TTL field in an IP packet.

When sending a packet via a shortcut, the ingress MPC must decrement the TTL by at least one in conformance with [Router Requirements]. There is no requirement to decrement the TTL by the actual number of router hops that are bypassed by the shortcut. If the TTL of the received packet is less than or equal to the value by which the MPC intends to decrement the TTL, the MPC must either send the packet unmodified to the MPS via LANE, or must discard the packet and send an ICMP Time Exceeded to the source of the packet.

Checksum

The IP checksum must be modified to reflect all changes to the IP header.

ICMP

The ingress MPC may encounter a variety of error conditions when forwarding packets and must ensure that the appropriate ICMP Error is generated. The ingress MPC must either send the packet unmodified to the MPS via LANE, or must send the applicable ICMP message.

An ingress MPC may generate the following ICMP messages:

- Destination Unreachable (Fragmentation needed but DF bit set)
- Time Exceeded (during transit)
- Parameter Problem

When sending the Destination Unreachable ICMP due to fragmentation, the Path MTU Discovery technique should be used as defined in [PATH MTU].

MTU

An ingress MPC may, but is not required to, fragment.

If the MTU returned in the MPOA Resolution Reply is smaller than the MTU on the inbound interface, the ingress MPC must make a decision on how to handle the shortcut and fragmentation. To avoid unnecessary packet mis-ordering, the ingress MPC must not set up the shortcut, and then send frames for a given flow on different paths based on whether fragmentation is required. The following are acceptable options:

1. Ingress MPC may establish the shortcut and fragment packets when necessary.
2. Ingress MPC may choose not to establish the shortcut.

A.1.5 Egress MPC Role

MTU

The egress MPC may advertise a large MTU and fragment the packets itself.

TTL

The egress MPC is not required to decrement TTL.

A.2 IPX Packet Handling in MPOA

A.2.1 Requirements

The MPOA System must support the IPX router requirements. MPOA distributes this responsibility across MPOA components.

A.2.2 Encapsulation

Four common IPX encapsulations are in use over Ethernet: Raw Ethernet (Novell proprietary), Ethernet_II, LLC, and LLC/SNAP. Other media types have their own associated encapsulations. Three IPX encapsulations, shown in Figure 20-Figure 22, may be used over a shortcut: Tagged, RFC 1483 Routed, and RFC 1483 NULL.

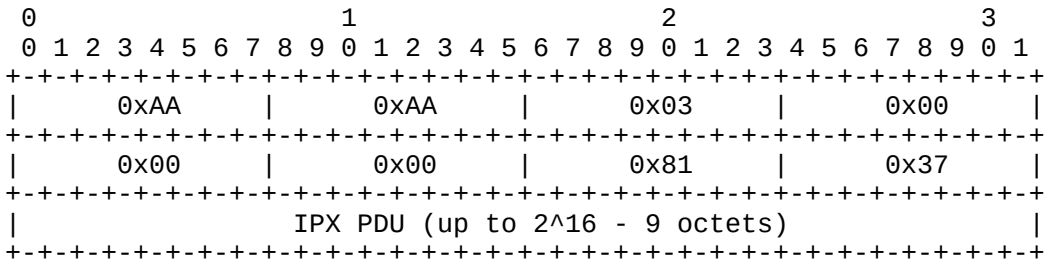


Figure 20 RFC 1483 LLC/SNAP Encapsulation for Routed IPX PDUs

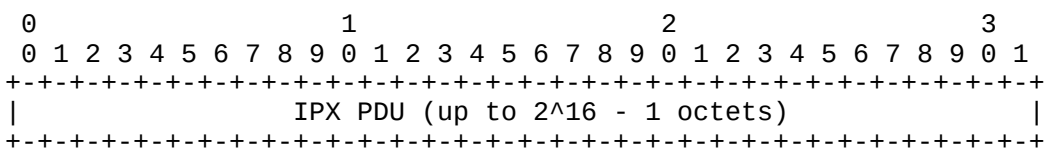


Figure 21 RFC 1483 “Null” Encapsulation for Routed IPX PDUs

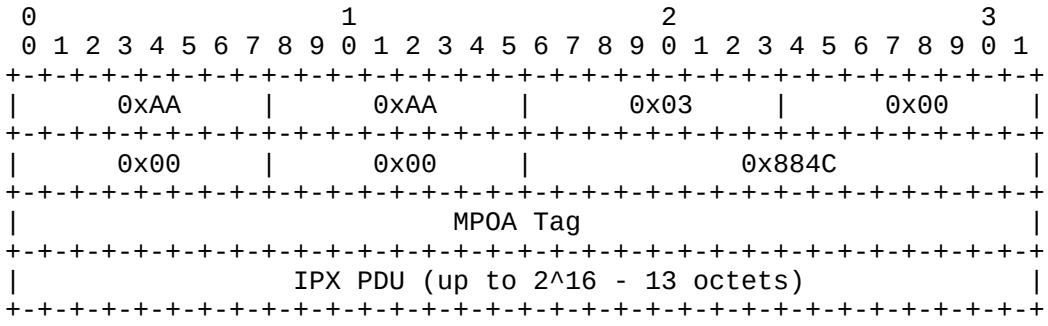


Figure 22 MPOA Tagged Encapsulation for IPX

A.2.3 MPS Role

An MPS is co-located with an IPX router and must process IPX packets in conformance with IPX router requirements.

The ingress MPS must include a hop count extension in MPOA Resolution Replies with a value as determined by the IPX routing protocol in use, (e.g. 16 for RIP).

A.2.4 Ingress MPC Role

IPX Options

There is no equivalent to the IP Options feature with IPX.

Transport Control

Unlike the IP TTL counts down, the IPX TC (Transport Control) field counts up. IPX hosts transmit packets with a TC of zero. When an upper bound is reached, the packet is discarded. An ingress MPC must increment the TC before sending an IPX packet on a shortcut.

Originally, the upper bound with the IPX-RIP routing protocol was 16. With NLSP, this number is configurable and may be as high as 128. Some other IPX routing protocols have an upper bound of 255. Packets that hit this limit may be dropped or forwarded unmodified to the router to be dropped. The checksum field is not updated when the TC is modified, as the TC field is not included in the set of fields the checksum covers.

An ingress MPC must only forward an IPX packet over a shortcut if the transport control field in the packet has a value that is less than the hop count extension value in the associated MPOA Resolution Reply. If the value of the transport control field is greater than or equal to the hop count extension value, the MPC must send the packet unmodified to the MPS via LANE.

In certain cases, unicast IPX packets may have a Source Network Number of zero. When a Netware client first transmits a packet after booting, it will not know its own network number and will insert a zero in the Source Network Number field. After the initial SAP/RIP exchange it will know its own network number and should use this from that point on. However, some clients do not do this and continue to use a zero Source Network Number. IPX routers are required to accept these packets and insert the correct Source Network Number as they forward a packet. If a packet with a source network number of zero is received by an MPC, the MPC must send the packet unmodified to the MPS via LANE.

MTU

There is no fragmentation in IPX. Netware applications try different packet sizes until communication is achieved.

Encapsulation

The ingress MPC removes the LAN encapsulation and adds the shortcut encapsulation.

A.2.5 Egress MPC Role

Checksum

When forwarding an IPX packet from a network where checksums are in use to a network using the Novell proprietary Raw Ethernet encapsulation, it is necessary to set the checksum field to 0xFFFF. Packets using the Raw Ethernet encapsulation can only be distinguished from other encapsulations if the checksum field is 0xFFFF. This processing must be carried out by the egress MPC.

Encapsulation

The egress MPC removes the shortcut encapsulation and adds the LAN encapsulation.

The egress MPC does no further processing of IPX packets and simply delivers the packets to the higher layers.

Annex B. MPOA Request/Reply Packet Contents [Normative]

B.1 Ingress MPC-Initiated MPOA Resolution

Ingress MPC-Initiated MPOA Resolution includes a request phase and a reply phase. The request phase proceeds from left to right as follows:

	Ingress MPC	**Ingress MPS**	**Egress MPS**	**Egress MPC**
Packet Type	MPOA Resolution Request	NHRP Resolution Request	MPOA Cache Imposition Request	
Request ID	Request ID 1	Request ID 2	Request ID 3	
Source Protocol Address	NULL or MPC Protocol Address	I-MPS Protocol Address	E-MPS Protocol Address	
Destination Protocol Address	Destination Protocol Address	Destination Protocol Address	Destination Protocol Address	
Source NBMA Address	I-MPC Data ATM Address	I-MPC Data ATM Address	I-MPC Data ATM Address	
Client Protocol Address (1)	NULL	NULL	NULL	
Prefix Length (1)	Widest Acceptable Prefix Length	Widest Acceptable Prefix Length	Requested Prefix Length	
Holding Time			≥ 2 * Holding Time	
Client NBMA Address (1)	NULL	NULL	NULL	
Extensions	Empty MPOA Egress Cache Tag Extension MPOA ATM Service Category Extension (1)	Received Extensions	Received Extensions MPOA DLL Header Extension (Cache ID, ELAN ID, DLL Header)	

The reply phase proceeds from right to left as follows:

	Ingress MPC	**Ingress MPS**	**Egress MPS**	**Egress MPC**
Packet Type		MPOA Resolution Reply	NHRP Resolution Reply	MPOA Cache Imposition Reply
Request ID		Request ID 1	Request ID 2	Request ID 3
Source Protocol Address		Restore to NULL or I-MPC address (from original MPOA Resolution Request)	I-MPS Protocol Address	E-MPS Protocol Address
Destination		Destination Protocol	Destination Protocol	Destination Protocol

Protocol Address		Address	Address	Address
Source NBMA Address		I-MPC Data ATM Address	I-MPC Data ATM Address	I-MPC Data ATM Address
Client Protocol Address		E-MPS Protocol Address	E-MPS Protocol Address	NULL
Prefix Length		Actual Prefix Length	Actual Prefix Length	Actual Prefix Length (2)
Holding Time		Holding Time	Holding Time	NULL
Client NBMA Address		E-MPC Data ATM Address	E-MPC Data ATM Address	E-MPC Data ATM Address
Extensions		Received Extensions	Received Extensions	Received Extensions

Legend:

NULL: zero length, no space allocated in packet

Notes:

(1) Optional

(2) An E-MPC can modify the Prefix Length to make it a host entry if a CIE was included in the request. An E-MPC must add a CIE with a host entry if a CIE was not included in the request.

B.2 Egress MPC-Initiated Egress Cache Purge

Egress MPC-Initiated Egress Cache Purge includes a request phase and a reply phase. The request phase proceeds from right to left as follows:

	Ingress MPC	**Ingress MPS**	**Egress MPS**	**Egress MPC**
Packet Type		NHRP Purge Request	NHRP Purge Request	MPOA Egress Cache Purge Request
Request ID		Request ID 3	Request ID 2	Request ID 1
Source Protocol Address		E-MPS Protocol Address	E-MPS Protocol Address	NULL
Destination Protocol Address		I-MPS Protocol Address	I-MPS Protocol Address	E-MPS Protocol Address
Source NBMA Address		E-MPC Data ATM Address	E-MPC Data ATM Address	E-MPC Data ATM Address
Client Protocol Address		Destination Protocol Address (to purge)	Destination Protocol Address (to purge)	Destination Protocol Address (to purge)
Prefix Length		Destination Prefix Length	Destination Prefix Length	Destination Prefix Length
Client NBMA		E-MPC Data ATM Address	E-MPC Data ATM Address	E-MPC Data ATM Address

Address				
Extensions		Received Extensions	Received Extensions	MPOA DLL Header Extension (Cache ID) Received Extensions

The reply phase proceeds from left to right as follows:

	Ingress MPC	**Ingress MPS**	**Egress MPS**	**Egress MPC**
Packet Type	NHRP Purge Reply	NHRP Purge Reply	MPOA Egress Cache Purge Reply	
Request ID	Request ID 3	Request ID 2	Request ID 1	
Source Protocol Address	E-MPS Protocol Address	E-MPS Protocol Address	NULL	
Destination Protocol Address	I-MPS Protocol Address	I-MPS Protocol Address	E-MPS Protocol Address	
Source NBMA Address	E-MPC Data ATM Address	E-MPC Data ATM Address	E-MPC Data ATM Address	
Client Protocol Address	Destination Protocol Address (to purge)	Destination Protocol Address (to purge)	Destination Protocol Address (to purge)	
Prefix Length	Destination Prefix Length	Destination Prefix Length	Destination Prefix Length	
Client NBMA Address	E-MPC Data ATM Address	E-MPC Data ATM Address	E-MPC Data ATM Address	
Extensions	Received Extensions	Received Extensions	Received Extensions	

B.3 Egress MPS-Initiated Egress Cache Purge

Egress MPS-Initiated Egress Cache Purges are transacted with the ingress MPC and the egress MPC simultaneously. Each transaction includes a request phase and a reply phase. The request phase proceeds as follows:

	Ingress MPC	**Ingress MPS**	**Egress MPS**	**Egress MPS**	**Egress MPC**
Packet Type		NHRP Purge Request	NHRP Purge Request	MPOA Cache Imposition Request	
Direction					
Request ID		Request ID 3	Request ID 2	Request ID 1	
Source Protocol Address		E-MPS Protocol Address	E-MPS Protocol Address	E-MPS Protocol Address	
Destination Protocol Address		I-MPS Addr	I-MPS Addr	Destination Protocol Address (to purge)	

Source NBMA Address		E-MPC Data ATM Address	E-MPC Data ATM Address	NULL	
Client Protocol Address		Destination Protocol Address (to purge)	Destination Protocol Address (to purge)	NULL	
Prefix Length		Destination Prefix Length	Destination Prefix Length	Destination Prefix Length	
Holding Time				0	
Client NBMA Address		E-MPC Data ATM Address	E-MPC Data ATM Address	NULL	
Extension				MPOA DLL Header Extension (Cache ID) (4)	

The reply phase proceeds as follows:

	Ingress MPC	**Ingress MPS**	**Egress MPS**	**Egress MPS**	**Egress MPC**
Packet Type	NHRP Purge Reply	NHRP Purge Reply			MPOA Cache Imposition Reply
Direction					
Request ID	Request ID 3	Request ID 2			Request ID 1
Source Protocol Address	E-MPS Protocol Address	E-MPS Protocol Address			E-MPS Protocol Address
Destination Protocol Address	I-MPS Protocol Address	I-MPS Protocol Address			Destination Protocol Address (to purge)
Source NBMA Address	E-MPC Data ATM Address	E-MPC Data ATM Address			NULL
Client Protocol Address	Destination Protocol Address (to purge)	Destination Protocol Address (to purge)			NULL
Prefix Length	Destination Prefix Length	Destination Prefix Length			Destination Prefix Length
Client NBMA Address	E-MPC Data ATM Address	E-MPC Data ATM Address			NULL
Extensions					Received Extensions

(4) Optional

B.4 Data-Plane Purge

Data-Plane Purges operate in a single phase from right to left as follows:

	Ingress MPC	Egress MPC
Packet Type		NHRP Purge Request
Request ID		Unused. Set to zero
Source Protocol Address		E-MPS Protocol Address or NULL (5)
Destination Protocol Address		NULL
Source NBMA Address		E-MPC Data ATM Address
Client Protocol Address		Destination Protocol Address (to purge)
Prefix Length		Destination Prefix Length
Client NBMA Address		E-MPC Data ATM Address
Extensions		

(5) Use E-MPS Protocol address if purge results from an MPS dying, and NULL if purge results from an egress cache miss.

B.5 MPOA Trigger

MPOA Triggers operate in a single phase from right to left as follows:

	Ingress MPC	Ingress MPS
Packet Type		MPOA Trigger
Request ID		unused
Source Protocol Address		NULL
Destination Protocol Address		NULL
Source NBMA Address		I-MPS Control ATM Address
Client Protocol Address		Destination Protocol Address (to trigger)
Prefix Length		Destination Prefix Length
Client NBMA Address		NULL
Extensions		None

B.6 MPOA Keep-Alive

MPOA Keep-Alive messages are sent by both ingress and egress MPSs. MPOA Keep-Alives operate in a single phase from left to right as follows:

	MPS	MPC
Packet Type	MPOA Keep-Alive	
Request ID	Keep-Alive sequence number	
Source Protocol Address	NULL	
Destination Protocol Address	NULL	
Source NBMA Address	MPS Control ATM Address	
Extensions	MPOA Keep-Alive Lifetime Extension	

Annex C. NBMA Next Hop Resolution Protocol (NHRP) [Normative]

This Annex contains a copy of the following Internet Draft:

[NHRP] Luciani, Katz, Piscitello, Cole, “NBMA Next Hop Resolution Protocol (NHRP).”, INTERNET DRAFT <draft-ietf-rolc-nhrp-11.txt>, expires September 1997.

It is the intent of the ATM Forum to replace this Annex with a reference to the official Request For Comments (RFC) for NHRP when it becomes available.

Routing over Large Clouds Working Group
INTERNET-DRAFT
<draft-ietf-rolc-nhrp-11.txt>

James V. Luciani
(Bay Networks)
Dave Katz
(cisco Systems)
David Piscitello
(Core Competence, Inc.)
Bruce Cole
(Juniper Networks)
Expires September 1997

NBMA Next Hop Resolution Protocol (NHRP)

Status of this Memo

Abstract

This document describes the NBMA Next Hop Resolution Protocol (NHRP). NHRP can be used by a source station (host or router) connected to a Non-Broadcast, Multi-Access (NBMA) subnetwork to determine the internetworking layer address and NBMA subnetwork addresses of the "NBMA next hop" towards a destination station. If the destination is connected to the NBMA subnetwork, then the NBMA next hop is the destination station itself. Otherwise, the NBMA next hop is the egress router from the NBMA subnetwork that is "nearest" to the destination station. NHRP is intended for use in a multiprotocol internetworking layer environment over NBMA subnetworks.

Note that while this protocol was developed for use with NBMA subnetworks, it is possible, if not likely, that it will be applied to BMA subnetworks as well. However, this usage of NHRP is for

further study.

This document is intended to be a functional superset of the NBMA Address Resolution Protocol (NARP) documented in [1].

Operation of NHRP as a means of establishing a transit path across an NBMA subnetwork between two routers will be addressed in a separate document (see [13]).

1. Introduction

The NBMA Next Hop Resolution Protocol (NHRP) allows a source station (a host or router), wishing to communicate over a Non-Broadcast, Multi-Access (NBMA) subnetwork, to determine the internetworking layer addresses and NBMA addresses of suitable "NBMA next hops" toward a destination station. A subnetwork can be non-broadcast either because it technically doesn't support broadcasting (e.g., an X.25 subnetwork) or because broadcasting is not feasible for one reason or another (e.g., an SMDS multicast group or an extended Ethernet would be too large). If the destination is connected to the NBMA subnetwork, then the NBMA next hop is the destination station itself. Otherwise, the NBMA next hop is the egress router from the NBMA subnetwork that is "nearest" to the destination station.

One way to model an NBMA network is by using the notion of logically independent IP subnets (LISs). LISs, as defined in [3] and [4], have the following properties:

1) All members of a LIS have the same IP network/subnet number and address mask.

2) All members of a LIS are directly connected to the same NBMA subnetwork.

3) All hosts and routers outside of the LIS are accessed via a router.

4) All members of a LIS access each other directly (without routers).

Address resolution as described in [3] and [4] only resolves the next hop address if the destination station is a member of the same LIS as the source station; otherwise, the source station must forward packets to a router that is a member of multiple LIS's. In multi-LIS configurations, hop-by-hop address resolution may not be sufficient to resolve the "NBMA next hop" toward the destination station, and IP packets may have multiple IP hops through the NBMA subnetwork.

Another way to model NBMA is by using the notion of Local Address

Groups (LAGs) [10]. The essential difference between the LIS and the LAG models is that while with the LIS model the outcome of the "local/remote" forwarding decision is driven purely by addressing information, with the LAG model the outcome of this decision is decoupled from the addressing information and is coupled with the Quality of Service and/or traffic characteristics. With the LAG model any two entities on a common NBMA network could establish a direct communication with each other, irrespective of the entities' addresses.

Support for the LAG model assumes the existence of a mechanism that allows any entity (i.e., host or router) connected to an NBMA network to resolve an internetworking layer address to an NBMA address for any other entity connected to the same NBMA network. This resolution would take place regardless of the address assignments to these entities. Within the parameters described in this document, NHRP describes such a mechanism. For example, when the internetworking layer address is of type IP, once the NBMA next hop has been resolved, the source may either start sending IP packets to the destination (in a connectionless NBMA subnetwork such as SMDS) or may first establish a connection to the destination with the desired bandwidth (in a connection-oriented NBMA subnetwork such as ATM).

Use of NHRP may be sufficient for hosts doing address resolution when those hosts are directly connected to an NBMA subnetwork, allowing for straightforward implementations in NBMA stations. NHRP also has the capability of determining the egress point from an NBMA subnetwork when the destination is not directly connected to the NBMA subnetwork and the identity of the egress router is not learned by other methods (such as routing protocols). Optional extensions to NHRP provide additional robustness and diagnosability.

Address resolution techniques such as those described in [3] and [4] may be in use when NHRP is deployed. ARP servers and services over NBMA subnetworks may be required to support hosts that are not capable of dealing with any model for communication other than the LIS model, and deployed hosts may not implement NHRP but may continue to support ARP variants such as those described in [3] and [4]. NHRP is intended to reduce or eliminate the extra router hops required by the LIS model, and can be deployed in a non-interfering manner with existing ARP services [14].

The operation of NHRP to establish transit paths across NBMA subnetworks between two routers requires additional mechanisms to avoid stable routing loops, and will be described in a separate document (see [13]).

2. Overview

2.1 Terminology

The term "network" is highly overloaded, and is especially confusing in the context of NHRP. We use the following terms:

Internetwork layer--the media-independent layer (IP in the case of TCP/IP networks).

Subnetwork layer--the media-dependent layer underlying the internetwork layer, including the NBMA technology (ATM, X.25, SMDS, etc.)

The term "server", unless explicitly stated to the contrary, refers to an Next Hop Server (NHS). An NHS is an entity performing the Next Hop Resolution Protocol service within the NBMA cloud. An NHS is always tightly coupled with a routing entity (router, route server or edge device) although the converse is not yet guaranteed until ubiquitous deployment of this functionality occurs. Note that the presence of intermediate routers that are not coupled with an NHS entity may preclude the use of NHRP when source and destination stations on different sides of such routers and thus such routers may partition NHRP reachability within an NBMA network.

The term "client", unless explicitly stated to the contrary, refers to an Next Hop Resolution Protocol client (NHC). An NHC is an entity which initiates NHRP requests of various types in order to obtain access to the NHRP service.

The term "station" generally refers to a host or router which contains an NHRP entity. Occasionally, the term station will describe a "user" of the NHRP client or service functionality; the difference in usage is largely semantic.

2.2 Protocol Overview

In this section, we briefly describe how a source S (which potentially can be either a router or a host) uses NHRP to determine the "NBMA next hop" to destination D.

For administrative and policy reasons, a physical NBMA subnetwork may be partitioned into several, disjoint "Logical NBMA subnetworks". A Logical NBMA subnetwork is defined as a collection of hosts and routers that share unfiltered subnetwork connectivity over an NBMA

subnetwork. "Unfiltered subnetwork connectivity" refers to the absence of closed user groups, address screening or similar features that may be used to prevent direct communication between stations connected to the same NBMA subnetwork. (Hereafter, unless otherwise specified, we use the term "NBMA subnetwork" to mean *logical* NBMA subnetwork.)

Placed within the NBMA subnetwork are one or more entities that implement the NHRP protocol. Such stations which are capable of answering NHRP Resolution Requests are known as "Next Hop Servers" (NHSs). Each NHS serves a set of destination hosts, which may or may not be directly connected to the NBMA subnetwork. NHSs cooperatively resolve the NBMA next hop within their logical NBMA subnetwork. In addition to NHRP, NHSs may support "classical" ARP service; however, this will be the subject of a separate document [14].

An NHS maintains a cache which contains protocol layer address to NBMA subnetwork layer address resolution information. This cache can be constructed from information obtained from NHRP Register packets (see Section 5.2.3 and 5.2.4), from NHRP Resolution Request/Reply packets, or through mechanisms outside the scope of this document (examples of such mechanisms might include ARP[3] and pre-configured tables). Section 6.2 further describes cache management issues.

For a station within a given LIS to avoid providing NHS functionality, there must be one or more NHSs within the NBMA subnetwork which are providing authoritative address resolution information on its behalf. Such an NHS is said to be "serving" the station. A stations on a LIS that lacks NHS functionality and is a client of the NHRP service is known as NHRP Client or just NHCs. If a serving NHS is to be able to supply the address resolution information for an NHC then NHSs must exist at each hop along all routed paths between the NHC making the resolution request and the destination NHC. The last NHRP entity along the routed path is the serving NHS; that is, NHRP Resolution Requests are not forwarded to destination NHCs but rather are processed by the serving NHS.

An NHC also maintains a cache of protocol address to NBMA address resolution information. This cache is populated through information obtained from NHRP Resolution Reply packets, from manual configuration, or through mechanisms outside the scope of this document.

The protocol proceeds as follows. An event occurs triggering station S to want to resolve the NBMA address of a path to D. This is most likely to be when a data packet addressed to station D is to be emitted from station S (either because station S is a host, or station S is a transit router), but the address resolution could also

be triggered by other means (a routing protocol update packet, for example). Station S first determines the next hop to station D through normal routing processes (for a host, the next hop may simply be the default router; for routers, this is the "next hop" to the destination internetwork layer address). If the destination's address resolution information is already available in S's cache then that information is used to forward the packet. Otherwise, if the next hop is reachable through one of its NBMA interfaces, S constructs an NHRP Resolution Request packet (see Section 5.2.1) containing station D's internetwork layer address as the (target) destination address, S's own internetwork layer address as the source address (Next Hop Resolution Request initiator), and station S's NBMA addressing information. Station S may also indicate that it prefers an authoritative NHRP Resolution Reply (i.e., station S only wishes to receive an NHRP Resolution Reply from an NHS serving the destination NHC). Station S emits the NHRP Resolution Request packet towards the destination.

If the NHRP Resolution Request is triggered by a data packet then S may, while awaiting an NHRP Resolution Reply, choose to dispose of the data packet in one of the following ways:

(a) Drop the packet
(b) Retain the packet until the NHRP Resolution Reply arrives and a more optimal path is available
(c) Forward the packet along the routed path toward D

The choice of which of the above to perform is a local policy matter, though option (c) is the recommended default, since it may allow data to flow to the destination while the NBMA address is being resolved. Note that an NHRP Resolution Request for a given destination MUST NOT be triggered on every packet.

When the NHS receives an NHRP Resolution Request, a check is made to see if it serves station D. If the NHS does not serve D, the NHS forwards the NHRP Resolution Request to another NHS. Mechanisms for determining how to forward the NHRP Resolution Request are discussed in Section 3.

If this NHS serves D, the NHS resolves station D's NBMA address information, and generates a positive NHRP Resolution Reply on D's behalf. NHRP Resolution Replies in this scenario are always marked as "authoritative". The NHRP Resolution Reply packet contains the address resolution information for station D which is to be sent back to S. Note that if station D is not on the NBMA subnetwork, the next hop internetwork layer address will be that of the egress router through which packets for station D are forwarded.

A transit NHS receiving an NHRP Resolution Reply may cache the address resolution information contained therein. To a subsequent NHRP Resolution Request, this NHS may respond with the cached, "non-authoritative" address resolution information if the NHS is permitted to do so (see Sections 5.2.2 and 6.2 for more information on non-authoritative versus authoritative NHRP Resolution Replies). Non-authoritative NHRP Resolution Replies are distinguished from authoritative NHRP Resolution Replies so that if a communication attempt based on non-authoritative information fails, a source station can choose to send an authoritative NHRP Resolution Request. NHSs MUST NOT respond to authoritative NHRP Resolution Requests with cached information.

If the determination is made that no NHS in the NBMA subnetwork can reply to the NHRP Resolution Request for D then a negative NHRP Resolution Reply (NAK) is returned. This occurs when (a) no next-hop resolution information is available for station D from any NHS, or (b) an NHS is unable to forward the NHRP Resolution Request (e.g., connectivity is lost).

NHRP Registration Requests, NHRP Purge Requests, NHRP Purge Replies, and NHRP Error Indications follow a routed path in the same fashion that NHRP Resolution Requests and NHRP Resolution Replies do. Specifically, "requests" and "indications" follow the routed path from Source Protocol Address (which is the address of the station initiating the communication) to the Destination Protocol Address. "Replies", on the other hand, follow the routed path from the Destination Protocol Address back to the Source Protocol Address with the following exceptions: in the case of a NHRP Registration Reply and in the case of an NHC initiated NHRP Purge Request, the packet is always returned via a direct VC (see Sections 5.2.4 and 5.2.5); if one does not exists then one MUST be created.

NHRP Requests and NHRP Replies do NOT cross the borders of a NBMA subnetwork however further study is being done in this area (see Section 7). Thus, the internetwork layer data traffic out of and into an NBMA subnetwork always traverses an internetwork layer router at its border.

NHRP optionally provides a mechanism to send a NHRP Resolution Reply which contains aggregated address resolution information. For example, suppose that router X is the next hop from station S to station D and that X is an egress router for all stations sharing an internetwork layer address prefix with station D. When an NHRP Resolution Reply is generated in response to a NHRP Resolution Request, the responder may augment the internetwork layer address of station D with a prefix length (see Section 5.2.0.1). A subsequent (non-authoritative) NHRP Resolution Request for some destination that

shares an internetwork layer address prefix (for the number of bits specified in the prefix length) with D may be satisfied with this cached information. See section 6.2 regarding caching issues.

To dynamically detect subnetwork-layer filtering in NBMA subnetworks (e.g., X.25 closed user group facility, or SMDS address screens), to trace the routed path that an NHRP packet takes, or to provide loop detection and diagnostic capabilities, a "Route Record" may be included in NHRP packets (see Sections 5.3.2 and 5.3.3). The Route Record extensions are the NHRP Forward Transit NHS Record Extension and the NHRP Reverse Transit NHS Record Extension. They contain the internetwork (and subnetwork layer) addresses of all intermediate NHSs between source and destination and between destination and source respectively. When a source station is unable to communicate with the responder (e.g., an attempt to open an SVC fails), it may attempt to do so successively with other subnetwork layer addresses in the NHRP Forward Transit NHS Record Extension until it succeeds (if authentication policy permits such action). This approach can find a suitable egress point in the presence of subnetwork-layer filtering (which may be source/destination sensitive, for instance, without necessarily creating separate logical NBMA subnetworks) or subnetwork-layer congestion (especially in connection-oriented media).

3. Deployment

NHRP Resolution Requests traverse one or more hops within an NBMA subnetwork before reaching the station that is expected to generate a response. Each station, including the source station, chooses a neighboring NHS to which it will forward the NHRP Resolution Request. The NHS selection procedure typically involves applying a destination protocol layer address to the protocol layer routing table which causes a routing decision to be returned. This routing decision is then used to forward the NHRP Resolution Request to the downstream NHS. The destination protocol layer address previously mentioned is carried within the NHRP Resolution Request packet. Note that even though a protocol layer address was used to acquire a routing decision, NHRP packets are not encapsulated within a protocol layer header but rather are carried at the NBMA layer using the encapsulation described in Section 5.

Each NHS/router examines the NHRP Resolution Request packet on its way toward the destination. Each NHS which the NHRP packet traverses on the way to the packet’s destination might modify the packet (e.g., updating the Forward Record extension). Ignoring error situations, the NHRP Resolution Request eventually arrives at a station that is to generate an NHRP Resolution Reply. This responding station

"serves" the destination. The responding station generates an NHRP Resolution Reply using the source protocol address from within the NHRP packet to determine where the NHRP Resolution Reply should be sent.

Rather than use routing to determine the next hop for an NHRP packet, an NHS may use other applicable means (such as static configuration information) in order to determine to which neighboring NHSs to forward the NHRP Resolution Request packet as long as such other means would not cause the NHRP packet to arrive at an NHS which is not along the routed path. The use of static configuration information for this purpose is beyond the scope of this document.

The NHS serving a particular destination must lie along the routed path to that destination. In practice, this means that all egress routers must double as NHSs serving the destinations beyond them, and that hosts on the NBMA subnetwork are served by routers that double as NHSs. Also, this implies that forwarding of NHRP packets within an NBMA subnetwork requires a contiguous deployment of NHRP capable routers. It is important that, in a given LIS/LAG which is using NHRP, all NHSs within the LIS/LAG have at least some portion of their resolution databases synchronized so that a packet arriving at one router/NHS in a given LIS/LAG will be forwarded in the same fashion as packet arriving at a different router/NHS for the given LIS/LAG. One method, among others, is to use the Server Cache Synchronization Protocol (SCSP) [12]. It is RECOMMENDED that SCSP be the method used when a LIS/LAG contains two or more router/NHSs.

During migration to NHRP, it cannot be expected that all routers within the NBMA subnetwork are NHRP capable. Thus, NHRP traffic which would otherwise need to be forwarded through such routers can be expected to be dropped due to the NHRP packet not being recognized. In this case, NHRP will be unable to establish any transit paths whose discovery requires the traversal of the non-NHRP speaking routers. If the client has tried and failed to acquire a cut through path then the client should use the network layer routed path as a default.

If an NBMA technology offers a group, an anycast, or a multicast addressing feature then the NHC may be configured with such an address which would be assigned to the appropriate NHSs. The NHC might then submit NHRP Resolution Requests to such an address, eliciting a response from one or more NHSs, depending on the response strategy selected. Note that the constraints described in Section 2 regarding directly sending NHRP Resolution Reply may apply.

When an NHS "serves" an NHC, the NHS MUST send NHRP messages destined

for the NHC directly to the NHC. That is, the NHRP message MUST NOT transit through any NHS which is not serving the NHC when the NHRP message is currently at an NHS which does serve the NHC (this, of course, assumes the NHRP message is destined for the NHC). Further, an NHS which serves an NHC SHOULD have a direct NBMA level connection to that NHC (see Section 5.2.3 and 5.2.4 for examples).

With the exception of NHRP Registration Requests (see Section 5.2.3 and 5.2.4 for details of the NHRP Registration Request case), an NHC MUST send NHRP messages over a direct NBMA level connection between the serving NHS and the served NHC.

It may not be desirable to maintain semi-permanent NBMA level connectivity between the NHC and the NHS. In this case, when NBMA level connectivity is initially setup between the NHS and the NHC (as described in Section 5.2.4), the NBMA address of the NHS should be obtained through the NBMA level signaling technology. This address should be stored for future use in setting up subsequent NBMA level connections. A somewhat more information rich technique to obtain the address information (and more) of the serving NHS would be for the NHC to include the Responder Address extension (see Section 5.3.1) in the NHRP Registration Request and to store the information returned to the NHC in the Responder Address extension which is subsequently included in the NHRP Registration Reply. Note also that, in practice, a client's default router should also be its NHS; thus a client may be able to know the NBMA address of its NHS from the configuration which was already required for the client to be able to communicate. Further, as mentioned in Section 4, NHCs may be configured with the addressing information of one or more NHSs.

4. Configuration

Next Hop Clients

An NHC connected to an NBMA subnetwork MAY be configured with the Protocol address(es) and NBMA address(es) of its NHS(s). The NHS(s) will likely also represent the NHC's default or peer routers, so their NBMA addresses may be obtained from the NHC's existing configuration. If the NHC is attached to several subnetworks (including logical NBMA subnetworks), the NHC should also be configured to receive routing information from its NHS(s) and peer routers so that it can determine which internetwork layer networks are reachable through which subnetworks.

Next Hop Servers

An NHS is configured with knowledge of its own internetwork layer and NBMA addresses. An NHS MAY also be configured with a set of internetwork layer address prefixes that correspond to the internetwork layer addresses of the stations it serves. The NBMA addresses of the stations served by the NHS may be learned via NHRP Registration packets.

If a served NHC is attached to several subnetworks, the router/route-server coresident with the serving NHS may also need to be configured to advertise routing information to such NHCs.

If an NHS acts as an egress router for stations connected to other subnetworks than the NBMA subnetwork, the NHS must, in addition to the above, be configured to exchange routing information between the NBMA subnetwork and these other subnetworks.

In all cases, routing information is exchanged using conventional intra-domain and/or inter-domain routing protocols.

5. NHRP Packet Formats

This section describes the format of NHRP packets. In the following, unless otherwise stated explicitly, the unqualified term "request" refers generically to any of the NHRP packet types which are "requests". Further, unless otherwise stated explicitly, the unqualified term "reply" refers generically to any of the NHRP packet types which are "replies".

An NHRP packet consists of a Fixed Part, a Mandatory Part, and an Extensions Part. The Fixed Part is common to all NHRP packet types. The Mandatory Part MUST be present, but varies depending on packet type. The Extensions Part also varies depending on packet type, and need not be present.

The length of the Fixed Part is fixed at 20 octets. The length of the Mandatory Part is determined by the contents of the extensions offset field (ar$extoff). If ar$extoff=0x0 then the mandatory part length is equal to total packet length (ar$pktsz) minus 20 otherwise the mandatory part length is equal to ar$extoff minus 20. The length of the Extensions Part is implied by ar$pktsz minus ar$extoff. NHSs may increase the size of an NHRP packet as a result of extension processing, but not beyond the offered maximum SDU size of the NBMA network.

NHRP packets are actually members of a wider class of address mapping

and management protocols being developed by the IETF. A specific encapsulation, based on the native formats used on the particular NBMA network over which NHRP is carried, indicates the generic IETF mapping and management protocol. For example, SMDS networks always use LLC/SNAP encapsulation at the NBMA layer [4], and an NHRP packet is preceded by the following LLC/SNAP encapsulation:

[0xAA-AA-03] [0x00-00-5E] [0x00-03]

The first three octets are LLC, indicating that SNAP follows. The SNAP OUI portion is the IANA's OUI, and the SNAP PID portion identifies the mapping and management protocol. A field in the Fixed Header following the encapsulation indicates that it is NHRP.

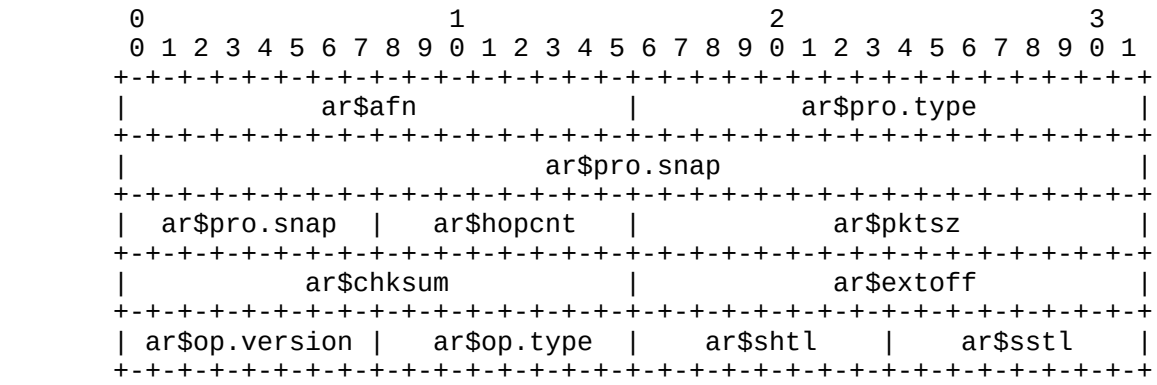
ATM uses either LLC/SNAP encapsulation of each packet (including NHRP), or uses no encapsulation on VCs dedicated to a single protocol (see [7]). Frame Relay and X.25 both use NLPID/SNAP encapsulation or identification of NHRP, using a NLPID of 0x0080 and the same SNAP contents as above (see [8], [9]).

Fields marked "unused" MUST be set to zero on transmission, and ignored on receipt.

Most packet types (ar$op.type) have both internetwork layer protocol-independent fields and protocol-specific fields. The protocol type/snap fields (ar$pro.type/snap) qualify the format of the protocol-specific fields.

5.1 NHRP Fixed Header

The Fixed Part of the NHRP packet contains those elements of the NHRP packet which are always present and do not vary in size with the type of packet.



ar$afn
Defines the type of "link layer" addresses being carried. This number is taken from the ’address family number’ list specified in [6]. This field has implications to the coding of ar$shtl and ar$sstl as described below.

ar$pro.type
field is a 16 bit unsigned integer representing the following number space:

0x0000 to 0x00FF	Protocols defined by the equivalent NLPIDs.
0x0100 to 0x03FF	Reserved for future use by the IETF.
0x0400 to 0x04FF	Allocated for use by the ATM Forum.
0x0500 to 0x05FF	Experimental/Local use.
0x0600 to 0xFFFF	Protocols defined by the equivalent Ethertypes.

(based on the observations that valid Ethertypes are never smaller than 0x600, and NLPIDs never larger than 0xFF.)

ar$pro.snap
When ar$pro.type has a value of 0x0080, a SNAP encoded extension is being used to encode the protocol type. This snap extension is placed in the ar$pro.snap field. This is termed the ’long form’ protocol ID. If ar$pro != 0x0080 then the ar$pro.snap field MUST be zero on transmit and ignored on receive. The ar$pro.type field itself identifies the protocol being referred to. This is termed the ’short form’ protocol ID.

In all cases, where a protocol has an assigned number in the ar$pro.type space (excluding 0x0080) the short form MUST be used when transmitting NHRP messages; i.e., if Ethertype or NLPID codings exist then they are used on transmit rather than the

ethertype. If both Ethertype and NLPID codings exist then when transmitting NHRP messages, the Ethertype coding MUST be used (this is consistent with RFC 1483 coding). So, for example, the following codings exist for IP:

```
SNAP:      ar$pro.type = 0x00-80, ar$pro.snap = 0x00-00-00-08-00
NLPID:     ar$pro.type = 0x00-CC, ar$pro.snap = 0x00-00-00-00-00
Ethertype: ar$pro.type = 0x08-00, ar$pro.snap = 0x00-00-00-00-00
```

and thus, since the Ethertype coding exists, it is used in preference.

ar$hopcnt
The Hop count indicates the maximum number of NHSs that an NHRP packet is allowed to traverse before being discarded. This field is used in a similar fashion to the way that a TTL is used in an IP packet and should be set accordingly. Each NHS decrements the TTL as the NHRP packet transits the NHS on the way to the next hop along the routed path to the destination. If an NHS receives an NHRP packet which it would normally forward to a next hop and that packet contains an ar$hopcnt set to zero then the NHS sends an error indication message back to the source protocol address stating that the hop count has been exceeded (see Section 5.2.7) and the NHS drops the packet in error; however, an error indication is never sent as a result of receiving an error indication. When a responding NHS replies to an NHRP request, that NHS places a value in ar$hopcnt as if it were sending a request of its own.

ar$pktsz
The total length of the NHRP packet, in octets (excluding link layer encapsulation).

ar$chksum
The standard IP checksum over the entire NHRP packet (starting with the fixed header). If only the hop count field is changed, the checksum is adjusted without full recomputation. The checksum is completely recomputed when other header fields are changed.

ar$extoff
This field identifies the existence and location of NHRP extensions. If this field is 0 then no extensions exist otherwise this field represents the offset from the beginning of the NHRP packet (i.e., starting from the ar$afn field) of the first extension.

ar$op.version
This field indicates what version of generic address mapping and

management protocol is represented by this message.

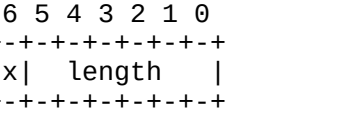
0	MARS protocol [11].
1	NHRP as defined in this document.
0x02 - 0xEF	Reserved for future use by the IETF.
0xF0 - 0xFE	Allocated for use by the ATM Forum.
0xFF	Experimental/Local use.

ar$op.type
When ar$op.version == 1, this is the NHRP packet type: NHRP Resolution Request(1), NHRP Resolution Reply(2), NHRP Registration Request(3), NHRP Registration Reply(4), NHRP Purge Request(5), NHRP Purge Reply(6), or NHRP Error Indication(7). Use of NHRP packet Types in the range 128 to 255 are reserved for research or use in other protocol development and will be administered by IANA.

ar$shtl
Type & length of source NBMA address interpreted in the context of the 'address family number'[6] indicated by ar$afn. See below for more details.

ar$sstl
Type & length of source NBMA subaddress interpreted in the context of the 'address family number'[6] indicated by ar$afn. When an NBMA technology has no concept of a subaddress, the subaddress length is always coded ar$sstl = 0 and no storage is allocated for the subaddress in the appropriate mandatory part. See below for more details.

Subnetwork layer address type/length fields (e.g., ar$shtl, Cli Addr T/L) and subnetwork layer subaddresses type/length fields (e.g., ar$sstl, Cli SAddr T/L) are coded as follows:



The most significant bit is reserved and MUST be set to zero. The second most significant bit (x) is a flag indicating whether the address being referred to is in:

- NSAP format (x = 0).
- Native E.164 format (x = 1).

For NBMA technologies that use neither NSAP nor E.164 format addresses, x = 0 SHALL be used to indicate the native form for the particular NBMA technology.

If the NBMA network is ATM and a subaddress (e.g., Source NBMA SubAddress, Client NBMA SubAddress) is to be included in any part of the NHRP packet then ar$afn MUST be set to 0x000F; further, the subnetwork layer address type/length fields (e.g., ar$shtl, Cli Addr T/L) and subnetwork layer subaddress type/length fields (e.g., ar$sstl, Cli SAddr T/L) MUST be coded as in [11]. If the NBMA network is ATM and no subaddress field is to be included in any part of the NHRP packet then ar$afn MAY be set to 0x0003 (NSAP) or 0x0008 (E.164) accordingly.

The bottom 6 bits is an unsigned integer value indicating the length of the associated NBMA address in octets. If this value is zero the flag x is ignored.

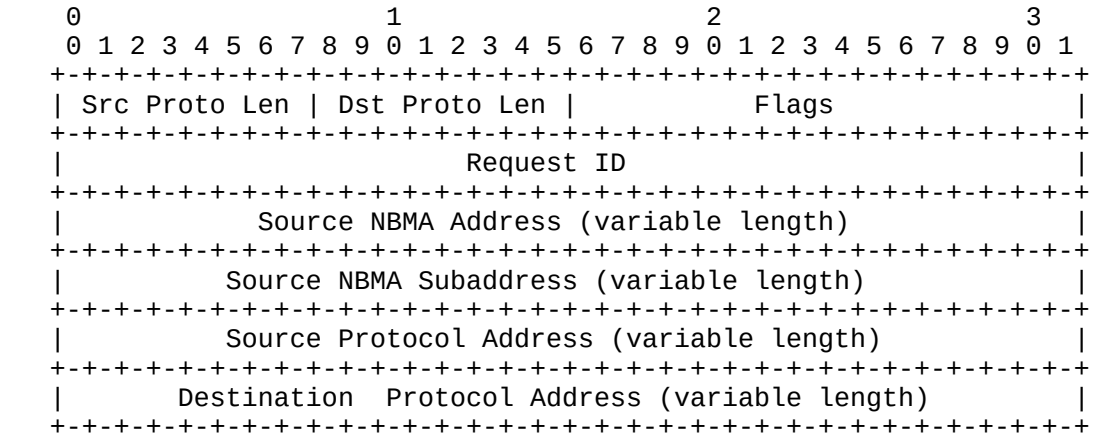
5.2.0 Mandatory Part

The Mandatory Part of the NHRP packet contains the operation specific information (e.g., NHRP Resolution Request/Reply, etc.) and variable length data which is pertinent to the packet type.

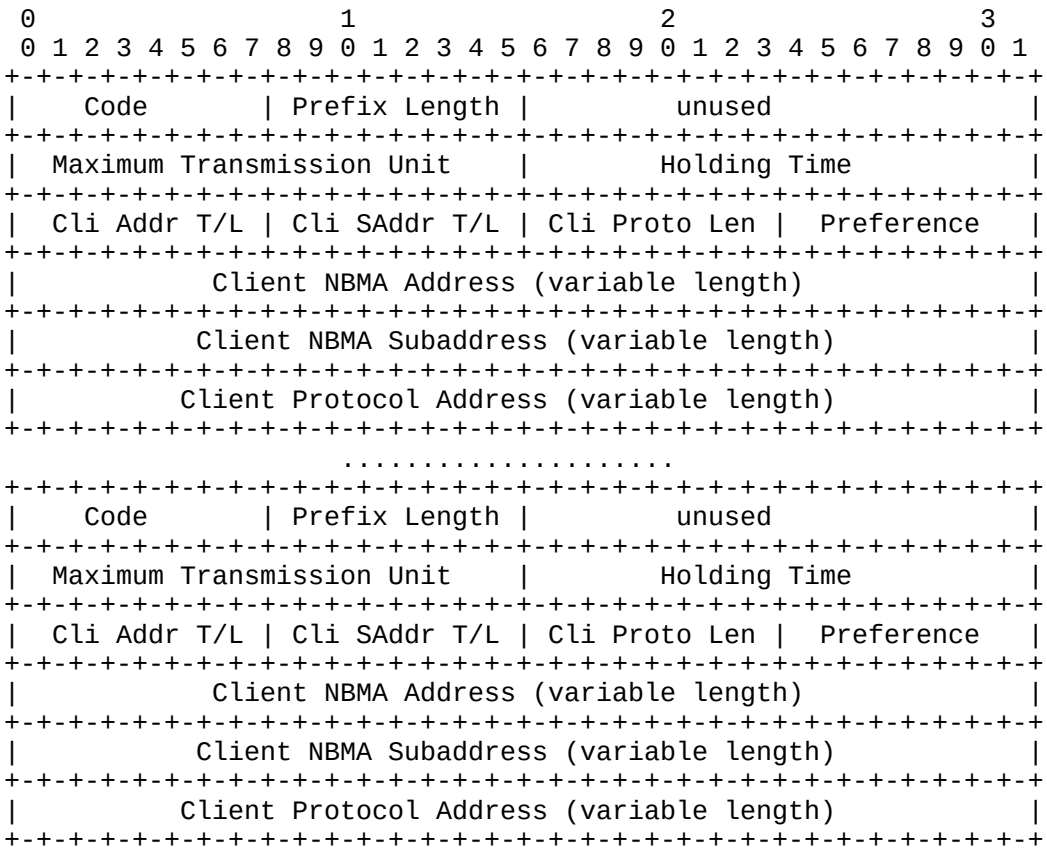
5.2.0.1 Mandatory Part Format

Sections 5.2.1 through 5.2.6 have a very similar mandatory part. This mandatory part includes a common header and zero or more Client Information Entries (CIEs). Section 5.2.7 has a different format which is specified in that section.

The common header looks like the following:



And the CIEs have the following format:



The meanings of the fields are as follows:

Src Proto Len
  This field holds the length in octets of the Source Protocol Address.

Dst Proto Len
  This field holds the length in octets of the Destination Protocol Address.

Flags
  These flags are specific to the given message type and they are explained in each section.

Request ID
  A value which, when coupled with the address of the source, provides a unique identifier for the information contained in a "request" packet. This value is copied directly from an "request"

packet into the associated "reply". When a sender of a "request" receives "reply", it will compare the Request ID and source address information in the received "reply" against that found in its outstanding "request" list. When a match is found then the "request" is considered to be acknowledged.

The value is taken from a 32 bit counter that is incremented each time a new "request" is transmitted. The same value MUST be used when resending a "request", i.e., when a "reply" has not been received for a "request" and a retry is sent after an appropriate interval.

The NBMA address/subaddress form specified below allows combined E.164/NSAPA form of NBMA addressing. For NBMA technologies without a subaddress concept, the subaddress field is always ZERO length and ar$sstl = 0.

Source NBMA Address

The Source NBMA address field is the address of the source station which is sending the "request". If the field's length as specified in ar$shtl is 0 then no storage is allocated for this address at all.

Source NBMA SubAddress

The Source NBMA subaddress field is the address of the source station which is sending the "request". If the field's length as specified in ar$sstl is 0 then no storage is allocated for this address at all.

For those NBMA technologies which have a notion of "Calling Party Addresses", the Source NBMA Addresses above are the addresses used when signaling for an SVC.

"Requests" and "indications" follow the routed path from Source Protocol Address to the Destination Protocol Address. "Replies", on the other hand, follow the routed path from the Destination Protocol Address back to the Source Protocol Address with the following exceptions: in the case of a NHRP Registration Reply and in the case of an NHC initiated NHRP Purge Request, the packet is always returned via a direct VC (see Sections 5.2.4 and 5.2.5).

Source Protocol Address

This is the protocol address of the station which is sending the "request". This is also the protocol address of the station toward which a "reply" packet is sent.

Destination Protocol Address

This is the protocol address of the station toward which a

"request" packet is sent.

Code
This field is message specific. See the relevant message sections below. In general, this field is a NAK code; i.e., when the field is 0 in a reply then the packet is acknowledging a request and if it contains any other value the packet contains a negative acknowledgment.

Prefix Length
This field is message specific. See the relevant message sections below. In general, however, this fields is used to indicate that the information carried in an NHRP message pertains to an equivalence class of internetwork layer addresses rather than just a single internetwork layer address specified. All internetwork layer addresses that match the first "Prefix Length" bit positions for the specific internetwork layer address are included in the equivalence class. If this field is set to 0x00 then this field MUST be ignored and no equivalence information is assumed (note that 0x00 is thus equivalent to 0xFF).

Maximum Transmission Unit
This field gives the maximum transmission unit for the relevant client station. If this value is 0 then either the default MTU is used or the MTU negotiated via signaling is used if such negotiation is possible for the given NBMA.

Holding Time
The Holding Time field specifies the number of seconds for which the Next Hop NBMA information specified in the CIE is considered to be valid. Cached information SHALL be discarded when the holding time expires. This field must be set to 0 on a NAK.

Cli Addr T/L
Type & length of next hop NBMA address specified in the CIE. This field is interpreted in the context of the ’address family number’[6] indicated by ar$afn (e.g., ar$afn=0x0003 for ATM).

Cli SAddr T/L
Type & length of next hop NBMA subaddress specified in the CIE. This field is interpreted in the context of the ’address family number’[6] indicated by ar$afn (e.g., ar$afn=0x0015 for ATM makes the address an E.164 and the subaddress an ATM Forum NSAP address). When an NBMA technology has no concept of a subaddress, the subaddress is always null with a length of 0. When the address length is specified as 0 no storage is allocated for the address.

Cli Proto Len
This field holds the length in octets of the Client Protocol Address specified in the CIE.

Preference
This field specifies the preference for use of the specific CIE relative to other CIEs. Higher values indicate higher preference. Action taken when multiple CIEs have equal or highest preference value is a local matter.

Client NBMA Address
This is the client's NBMA address.

Client NBMA SubAddress
This is the client's NBMA subaddress.

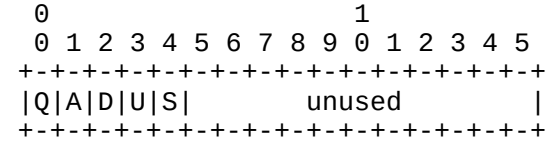
Client Protocol Address
This is the client's internetworking layer address specified.

Note that an NHS may cache source address binding information from an NHRP Resolution Request if and only if the conditions described in Section 6.2 are met for the NHS. In all other cases, source address binding information appearing in an NHRP message MUST NOT be cached.

5.2.1 NHRP Resolution Request

The NHRP Resolution Request packet has a Type code of 1. Its mandatory part is coded as described in Section 5.2.0.1 and the message specific meanings of the fields are as follows:

Flags - The flags field is coded as follows:



Q
Set if the station sending the NHRP Resolution Request is a router; clear if the it is a host.

A
This bit is set in a NHRP Resolution Request if only authoritative next hop information is desired and is clear otherwise. See the NHRP Resolution Reply section below for further details on the "A" bit and its usage.

D
Unused (clear on transmit)

U
This is the Uniqueness bit. This bit aids in duplicate address detection. When this bit is set in an NHRP Resolution Request and one or more entries exist in the NHS cache which meet the requirements of the NHRP Resolution Request then only the CIE in the NHS's cache with this bit set will be returned. Note that even if this bit was set at registration time, there may still be multiple CIEs that might fulfill the NHRP Resolution Request because an entire subnet can be registered through use of the Prefix Length in the CIE and the address of interest might be within such a subnet. If the "uniqueness" bit is set and the responding NHS has one or more cache entries which match the request but no such cache entry has the "uniqueness" bit set, then the NHRP Resolution Reply returns with a NAK code of "13 - Binding Exists But Is Not Unique" and no CIE is included. If a client wishes to receive non- unique Next Hop Entries, then the client must have the "uniqueness" bit set to zero in its NHRP Resolution Request. Note that when this bit is set in an NHRP Registration Request, only a single CIE may be specified in the NHRP Registration Request and that CIE must have the Prefix Length field set to 0xFF.

S
Set if the binding between the Source Protocol Address and the Source NBMA information in the NHRP Resolution Request is guaranteed to be stable and accurate (e.g., these addresses are those of an ingress router which is connected to an ethernet stub network or the NHC is an NBMA attached host).

Zero or one CIEs (see Section 5.2.0.1) may be specified in an NHRP Resolution Request. If one is specified then that entry carries the pertinent information for the client sourcing the NHRP Resolution Request. Usage of the CIE in the NHRP Resolution Request is described below:

Prefix Length
If a CIE is specified in the NHRP Resolution Request then the Prefix Length field may be used to qualify the widest acceptable prefix which may be used to satisfy the NHRP Resolution Request. In the case of NHRP Resolution Request/Reply, the Prefix Length specifies the equivalence class of addresses which match the first "Prefix Length" bit positions of the Destination Protocol Address. If the "U" bit is set in the common header then this field MUST be set to 0xFF.

Maximum Transmission Unit
This field gives the maximum transmission unit for the source station. A possible use of this field in the NHRP Resolution Request packet is for the NHRP Resolution Requester to ask for a target MTU. In lieu of that usage, the CIE must be omitted.

Holding Time
The Holding Time specified in the one CIE permitted to be included in an NHRP Resolution Request is the amount of time which the source address binding information in the NHRP Resolution Request is permitted to cached by transit and responding NHSs. Note that this field may only have a non-zero value if the S bit is set.

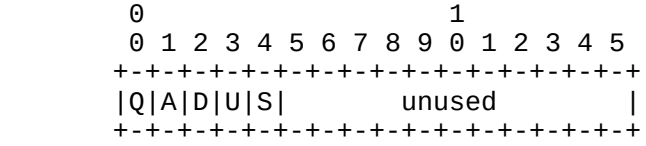
All other fields in the CIE MUST be ignored and SHOULD be set to 0.

The Destination Protocol Address in the common header of the Mandatory Part of this message contains the protocol address of the station for which resolution is desired. An NHC MUST send the NHRP Resolution Request directly to one of its serving NHSs (see Section 3 for more information).

5.2.2 NHRP Resolution Reply

The NHRP Resolution Reply packet has a Type code of 2. CIEs correspond to Next Hop Entries in an NHS's cache which match the criteria in the NHRP Resolution Request. Its mandatory part is coded as described in Section 5.2.0.1. The message specific meanings of the fields are as follows:

Flags - The flags field is coded as follows:



Q
Copied from the NHRP Resolution Request. Set if the NHRP Resolution Requester is a router; clear if it is a host.

A
Set if the next hop CIE in the NHRP Resolution Reply is authoritative; clear if the NHRP Resolution Reply is non-authoritative.

When an NHS receives a NHRP Resolution Request for authoritative information for which it is the authoritative source, it MUST respond with a NHRP Resolution Reply containing all and only those next hop CIEs which are contained in the NHS's cache which both match the criteria of the NHRP Resolution Request and are authoritative cache entries. An NHS is an authoritative source for a NHRP Resolution Request if the information in the NHS's cache matches the NHRP Resolution Request criteria and that information was obtained through a NHRP Registration Request or through synchronization with an NHS which obtained this information through a NHRP Registration Request. An authoritative cache entry is one which is obtained through a NHRP Registration Request or through synchronization with an NHS which obtained this information through a NHRP Registration Request.

An NHS obtains non-authoritative CIEs through promiscuous listening to NHRP packets other than NHRP Registrations which are directed at it. A NHRP Resolution Request which indicates a request for non-authoritative information should cause a NHRP Resolution Reply which contains all entries in the replying NHS's cache (i.e., both authoritative and non-authoritative) which match the criteria specified in the request.

D

Set if the association between destination and the associate next hop information included in all CIEs of the NHRP Resolution Reply is guaranteed to be stable for the lifetime of the information (the holding time). This is the case if the Next Hop protocol address in a CIE identifies the destination (though it may be different in value than the Destination address if the destination system has multiple addresses) or if the destination is not connected directly to the NBMA subnetwork but the egress router to that destination is guaranteed to be stable (such as when the destination is immediately adjacent to the egress router through a non-NBMA interface).

U

This is the Uniqueness bit. See the NHRP Resolution Request section above for details. When this bit is set only, only one CIE is included since only one unique binding should exist in an NHS's cache.

S

Copied from NHRP Resolution Request message.

One or more CIEs are specified in the NHRP Resolution Reply. Each CIE contains NHRP next hop information which the responding NHS has cached and which matches the parameters specified in the NHRP

Resolution Request. If no match is found by the NHS issuing the NHRP Resolution Reply then a single CIE is enclosed with the a CIE Code set appropriately (see below) and all other fields MUST be ignored and SHOULD be set to 0. In order to facilitate the use of NHRP by minimal client implementations, the first CIE MUST contain the next hop with the highest preference value so that such an implementation need parse only a single CIE.

Code

If this field is set to zero then this packet contains a positively acknowledged NHRP Resolution Reply. If this field contains any other value then this message contains an NHRP Resolution Reply NAK which means that an appropriate internetworking layer to NBMA address binding was not available in the responding NHS's cache. If NHRP Resolution Reply contains a Client Information Entry with a NAK Code other than 0 then it MUST NOT contain any other CIE. Currently defined NAK Codes are as follows:

4 - Administratively Prohibited

An NHS may refuse an NHRP Resolution Request attempt for administrative reasons (due to policy constraints or routing state). If so, the NHS MUST send an NHRP Resolution Reply which contains a NAK code of 4.

5 - Insufficient Resources

If an NHS cannot serve a station due to a lack of resources (e.g., can't store sufficient information to send a purge if routing changes), the NHS MUST reply with a NAKed NHRP Resolution Reply which contains a NAK code of 5.

12 - No Internetworking Layer Address to NBMA Address Binding Exists

This code states that there were absolutely no internetworking layer address to NBMA address bindings found in the responding NHS's cache.

13 - Binding Exists But Is Not Unique

This code states that there were one or more internetworking layer address to NBMA address bindings found in the responding NHS's cache, however none of them had the uniqueness bit set.

Prefix Length

In the case of NHRP Resolution Reply, the Prefix Length specifies the equivalence class of addresses which match the first "Prefix Length" bit positions of the Destination Protocol Address.

Holding Time
The Holding Time specified in a CIE of an NHRP Resolution Reply is the amount of time remaining before the expiration of the client information which is cached at the replying NHS. It is not the value which was registered by the client.

The remainder of the fields for the CIE for each next hop are filled out as they were defined when the next hop was registered with the responding NHS (or one of the responding NHS's synchronized servers) via the NHRP Registration Request.

Load-splitting may be performed when more than one Client Information Entry is returned to a requester when equal preference values are specified. Also, the alternative addresses may be used in case of connectivity failure in the NBMA subnetwork (such as a failed call attempt in connection-oriented NBMA subnetworks).

Any extensions present in the NHRP Resolution Request packet MUST be present in the NHRP Resolution Reply even if the extension is non-Compulsory.

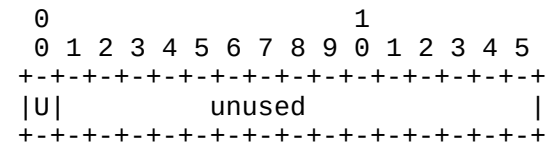
If an unsolicited NHRP Resolution Reply packet is received, an Error Indication of type Invalid NHRP Resolution Reply Received SHOULD be sent in response.

When an NHS that serves a given NHC receives an NHRP Resolution Reply destined for that NHC then the NHS must MUST send the NHRP Resolution Reply directly to the NHC (see Section 3).

5.2.3 NHRP Registration Request

The NHRP Registration Request is sent from a station to an NHS to notify the NHS of the station's NBMA information. It has a Type code of 3. Each CIE corresponds to Next Hop information which is to be cached at an NHS. The mandatory part of an NHRP Registration Request is coded as described in Section 5.2.0.1. The message specific meanings of the fields are as follows:

Flags - The flags field is coded as follows:



U
This is the Uniqueness bit. When set in an NHRP Registration Request, this bit indicates that the registration of the protocol address is unique within the confines of the set of synchronized NHSs. This "uniqueness" qualifier MUST be stored in the NHS/NHC cache. Any attempt to register a binding between the protocol address and an NBMA address when this bit is set MUST be rejected with a Code of "14 - Unique Internetworking Layer Address Already Registered" if the replying NHS already has a cache entry for the protocol address and the cache entry has the "uniqueness" bit set. A registration of a CIE's information is rejected when the CIE is returned with the Code field set to anything other than 0x00. See the description of the uniqueness bit in NHRP Resolution Request section above for further details. When this bit is set only, only one CIE MAY be included in the NHRP Registration Request.

Request ID
The request ID has the same meaning as described in Section 5.2.0.1. However, the request ID for NHRP Registrations which is maintained at each client MUST be kept in non-volatile memory so that when a client crashes and reregisters there will be no inconsistency in the NHS's database. In order to reduce the overhead associated with updating non-volatile memory, the actual updating need not be done with every increment of the Request ID but could be done, for example, every 50 or 100 increments. In this scenario, when a client crashes and reregisters it knows to add 100 to the value of the Request ID in the non-volatile memory before using the Request ID for subsequent registrations.

One or more CIEs are specified in the NHRP Registration Request. Each CIE contains next hop information which a client is attempting to register with its servers. Generally, all fields in CIEs enclosed in NHRP Registration Requests are coded as described in Section 5.2.0.1. However, if a station is only registering itself with the NHRP Registration Request then it MAY code the Cli Addr T/L, Cli SAddr T/L, and Cli Proto Len as zero which signifies that the client address information is to be taken from the source information in the common header (see Section 5.2.0.1). Below, further clarification is given for some fields in a CIE in the context of a NHRP Registration

Request.

Code
This field is set to 0x00 in NHRP Registration Requests.

Prefix Length

This field may be used in a NHRP Registration Request to register equivalence information for the Client Protocol Address specified in the CIE of an NHRP Registration Request In the case of NHRP Registration Request, the Prefix Length specifies the equivalence class of addresses which match the first "Prefix Length" bit positions of the Client Protocol Address. If the "U" bit is set in the common header then this field MUST be set to 0xFF.

The NHRP Registration Request is used to register an NHC's NHRP information with its NHSs. If an NHC is configured with the protocol address of a serving NHS then the NHC may place the NHS's protocol address in the Destination Protocol Address field of the NHRP Registration Request common header otherwise the NHC must place its own protocol address in the Destination Protocol Address field.

When an NHS receives an NHRP Registration Request which has the Destination Protocol Address field set to an address which belongs to a LIS/LAG for which the NHS is serving then if the Destination Protocol Address field is equal to the Source Protocol Address field (which would happen if the NHC put its protocol address in the Destination Protocol Address) or the Destination Protocol Address field is equal to the protocol address of the NHS then the NHS processes the NHRP Registration Request after doing appropriate error checking (including any applicable policy checking).

When an NHS receives an NHRP Registration Request which has the Destination Protocol Address field set to an address which does not belong to a LIS/LAG for which the NHS is serving then the NHS forwards the packet down the routed path toward the appropriate LIS/LAG.

When an NHS receives an NHRP Registration Request which has the Destination Protocol Address field set to an address which belongs to a LIS/LAG for which the NHS is serving then if the Destination Protocol Address field does not equal the Source Protocol Address field and the Destination Protocol Address field does not equal the protocol address of the NHS then the NHS forwards the message to the appropriate NHS within the LIS/LAG as specified by Destination Protocol Address field.

It is possible that a misconfigured station will attempt to register

with the wrong NHS (i.e., one that cannot serve it due to policy constraints or routing state). If this is the case, the NHS MUST reply with a NAK-ed Registration Reply of type Can't Serve This Address.

If an NHS cannot serve a station due to a lack of resources, the NHS MUST reply with a NAK-ed Registration Reply of type Registration Overflow.

In order to keep the registration entry from being discarded, the station MUST re-send the NHRP Registration Request packet often enough to refresh the registration, even in the face of occasional packet loss. It is recommended that the NHRP Registration Request packet be sent at an interval equal to one-third of the Holding Time specified therein.

5.2.4 NHRP Registration Reply

The NHRP Registration Reply is sent by an NHS to a client in response to that client's NHRP Registration Request. If the Code field of a CIE in the NHRP Registration Reply has anything other than 0 zero in it then the NHRP Registration Reply is a NAK otherwise the reply is an ACK. The NHRP Registration Reply has a Type code of 4.

An NHRP Registration Reply is formed from an NHRP Registration Request by changing the type code to 4, updating the CIE Code field, and filling in the appropriate extensions if they exist. The message specific meanings of the fields are as follows:

Attempts to register the information in the CIEs of an NHRP Registration Request may fail for various reasons. If this is the case then each failed attempt to register the information in a CIE of an NHRP Registration Request is logged in the associated NHRP Registration Reply by setting the CIE Code field to the appropriate error code as shown below:

CIE Code

0 - Successful Registration

The information in the CIE was successfully registered with the NHS.

4 - Administratively Prohibited

An NHS may refuse an NHRP Registration Request attempt for administrative reasons (due to policy constraints or routing

state). If so, the NHS MUST send an NHRP Registration Reply which contains a NAK code of 4.

5 - Insufficient Resources

If an NHS cannot serve a station due to a lack of resources, the NHS MUST reply with a NAKed NHRP Registration Reply which contains a NAK code of 5.

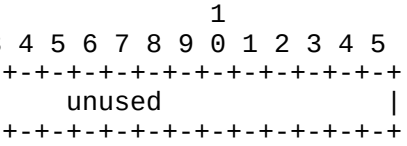
14 - Unique Internetworking Layer Address Already Registered
If a client tries to register a protocol address to NBMA address binding with the uniqueness bit on and the protocol address already exists in the NHS's cache then if that cache entry also has the uniqueness bit on then this NAK Code is returned in the CIE in the NHRP Registration Reply.

Due to the possible existence of asymmetric routing, an NHRP Registration Reply may not be able to merely follow the routed path back to the source protocol address specified in the common header of the NHRP Registration Reply. As a result, there MUST exist a direct NBMA level connection between the NHC and its NHS on which to send the NHRP Registration Reply before NHRP Registration Reply may be returned to the NHC. If such a connection does not exist then the NHS must setup such a connection to the NHC by using the source NBMA information supplied in the common header of the NHRP Registration Request.

5.2.5 NHRP Purge Request

The NHRP Purge Request packet is sent in order to invalidate cached information in a station. The NHRP Purge Request packet has a type code of 5. The mandatory part of an NHRP Purge Request is coded as described in Section 5.2.0.1. The message specific meanings of the fields are as follows:

Flags - The flags field is coded as follows:



N
When set, this bit tells the receiver of the NHRP Purge Request that the requester does not expect to receive an NHRP Purge Reply. If an unsolicited NHRP Purge Reply is received by a

station where that station is identified in the Source Protocol Address of the packet then that packet must be ignored.

One or more CIEs are specified in the NHRP Purge Request. Each CIE contains next hop information which is to be purged from an NHS/NHC cache. Generally, all fields in CIEs enclosed in NHRP Purge Requests are coded as described in Section 5.2.0.1. Below, further clarification is given for some fields in a CIE in the context of a NHRP Purge Request.

Code
This field is set to 0x00 in NHRP Purge Requests.

Prefix Length

In the case of NHRP Purge Requests, the Prefix Length specifies the equivalence class of addresses which match the first "Prefix Length" bit positions of the Client Protocol Address specified in the CIE. All next hop information which contains a protocol address which matches an element of this equivalence class is to be purged from the receivers cache.

The Maximum Transmission Unit and Preference fields of the CIE are coded as zero. The Holding Time should be coded as zero but there may be some utility in supplying a "short" holding time to be applied to the matching next hop information before that information would be purged; this usage is for further study. The Client Protocol Address field and the Cli Proto Len field MUST be filled in. The Client Protocol Address is filled in with the protocol address to be purged from the receiving station’s cache while the Cli Proto Len is set the length of the purged client’s protocol address. All remaining fields in the CIE MAY be set to zero although the client NBMA information (and associated length fields) MAY be specified to narrow the scope of the NHRP Purge Request if requester desires. However, the receiver of an NHRP Purge Request may choose to ignore the Client NBMA information if it is supplied.

An NHRP Purge Request packet is sent from an NHS to a station to cause it to delete previously cached information. This is done when the information may be no longer valid (typically when the NHS has previously provided next hop information for a station that is not directly connected to the NBMA subnetwork, and the egress point to that station may have changed).

An NHRP Purge Request packet may also be sent from an NHC to an NHS with which the NHC had previously registered. This allows for an NHC to invalidate its registration with NHRP before it would otherwise

expire via the holding timer. If an NHC does not have knowledge of a protocol address of a serving NHS then the NHC must place its own protocol address in the Destination Protocol Address field and forward the packet along the routed path. Otherwise, the NHC must place the protocol address of a serving NHS in this field.

Serving NHSs may need to send one or more new NHRP Purge Requests as a result of receiving a purge from one of their served NHCs since the NHS may have previously responded to NHRP Resolution Requests for that NHC's NBMA information. These purges are "new" in that they are sourced by the NHS and not the NHC; that is, for each NHC that previously sent a NHRP Resolution Request for the purged NHC NBMA information, an NHRP Purge Request is sent which contains the Source Protocol/NBMA Addresses of the NHS and the Destination Protocol Address of the NHC which previously sent an NHRP Resolution Request prior to the purge.

The station sending the NHRP Purge Request MAY periodically retransmit the NHRP Purge Request until either NHRP Purge Request is acknowledged or until the holding time of the information being purged has expired. Retransmission strategies for NHRP Purge Requests are a local matter.

When a station receives an NHRP Purge Request, it MUST discard any previously cached information that matches the information in the CIEs.

An NHRP Purge Reply MUST be returned for the NHRP Purge Request even if the station does not have a matching cache entry assuming that the "N" bit is off in the NHRP Purge Request.

If the station wishes to reestablish communication with the destination shortly after receiving an NHRP Purge Request, it should make an authoritative NHRP Resolution Request in order to avoid any stale cache entries that might be present in intermediate NHSs (See section 6.2.2.). It is recommended that authoritative NHRP Resolution Requests be made for the duration of the holding time of the old information.

5.2.6 NHRP Purge Reply

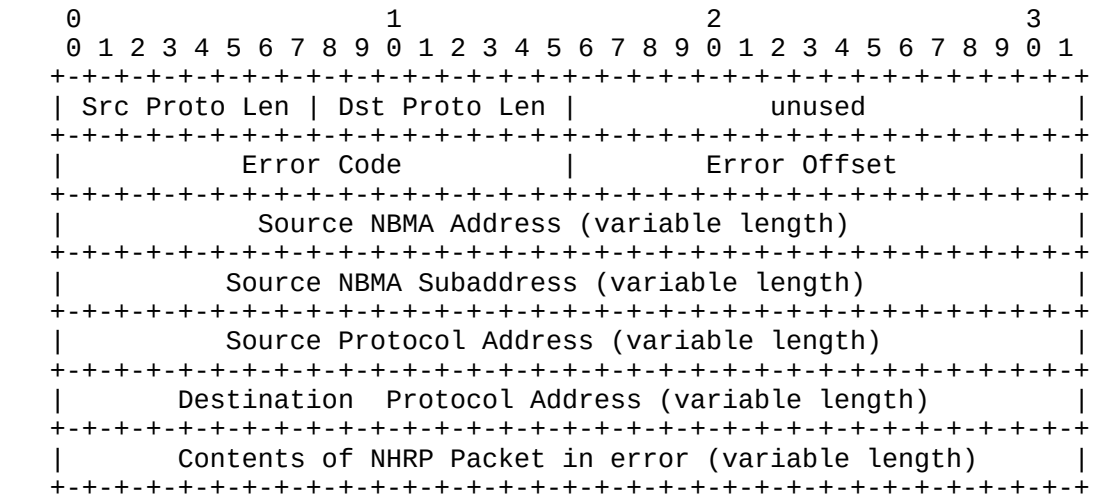
The NHRP Purge Reply packet is sent in order to assure the sender of an NHRP Purge Request that all cached information of the specified type has been purged from the station sending the reply. The NHRP Purge Reply has a type code of 6.

An NHRP Purge Reply is formed from an NHRP Purge Request by merely

changing the type code in the request to 6. The packet is then returned to the requester after filling in the appropriate extensions if they exist.

5.2.7 NHRP Error Indication

The NHRP Error Indication is used to convey error indications to the sender of an NHRP packet. It has a type code of 7. The Mandatory Part has the following format:



Src Proto Len
This field holds the length in octets of the Source Protocol Address.

Dst Proto Len
This field holds the length in octets of the Destination Protocol Address.

Error Code
An error code indicating the type of error detected, chosen from the following list:

1 - Unrecognized Extension

When the Compulsory bit of an extension in NHRP packet is set, the NHRP packet cannot be processed unless the extension has been processed. The responder MUST return an NHRP Error Indication of type Unrecognized Extension if it is incapable of

processing the extension. However, if a transit NHS (one which is not going to generate a reply) detects an unrecognized extension, it SHALL ignore the extension.

3 - NHRP Loop Detected

A Loop Detected error is generated when it is determined that an NHRP packet is being forwarded in a loop.

6 - Protocol Address Unreachable

This error occurs when a packet it moving along the routed path and it reaches a point such that the protocol address of interest is not reachable.

7 - Protocol Error

A generic packet processing error has occurred (e.g., invalid version number, invalid protocol type, failed checksum, etc.)

8 - NHRP SDU Size Exceeded

If the SDU size of the NHRP packet exceeds the MTU size of the NBMA network then this error is returned.

9 - Invalid Extension

If an NHS finds an extension in a packet which is inappropriate for the packet type, an error is sent back to the sender with Invalid Extension as the code.

10 - Invalid NHRP Resolution Reply Received

If a client receives a NHRP Resolution Reply for a Next Hop Resolution Request which it believes it did not make then an error packet is sent to the station making the reply with an error code of Invalid Reply Received.

11 - Authentication Failure

If a received packet fails an authentication test then this error is returned.

15 - Hop Count Exceeded

The hop count which was specified in the Fixed Header of an NHRP message has been exceeded.

Error Offset
The offset in octets into the NHRP packet, starting at the NHRP Fixed Header, at which the error was detected.

Source NBMA Address
The Source NBMA address field is the address of the station which observed the error.

Source NBMA SubAddress
The Source NBMA subaddress field is the address of the station which observed the error. If the field's length as specified in ar$sstl is 0 then no storage is allocated for this address at all.

Source Protocol Address
This is the protocol address of the station which issued the Error packet.

Destination Protocol Address
This is the protocol address of the station which sent the packet which was found to be in error.

An NHRP Error Indication packet SHALL NEVER be generated in response to another NHRP Error Indication packet. When an NHRP Error Indication packet is generated, the offending NHRP packet SHALL be discarded. In no case should more than one NHRP Error Indication packet be generated for a single NHRP packet.

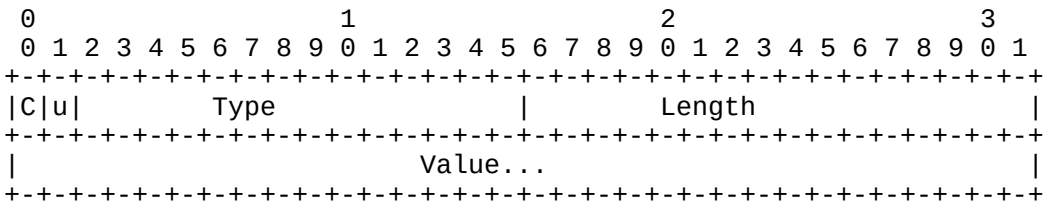
If an NHS sees its own Protocol and NBMA Addresses in the Source NBMA and Source Protocol address fields of a transiting NHRP Error Indication packet then the NHS will quietly drop the packet and do nothing (this scenario would occur when the NHRP Error Indication packet was itself in a loop).

Note that no extensions may be added to an NHRP Error Indication.

5.3 Extensions Part

The Extensions Part, if present, carries one or more extensions in {Type, Length, Value} triplets.

Extensions have the following format:



C
"Compulsory." If clear, and the NHS does not recognize the type code, the extension may safely be ignored. If set, and the NHS does not recognize the type code, the NHRP "request" is considered to be in error. (See below for details.)

u
Unused and must be set to zero.

Type
The extension type code (see below). The extension type is not qualified by the Compulsory bit, but is orthogonal to it.

Length
The length in octets of the value (not including the Type and Length fields; a null extension will have only an extension header and a length of zero).

When extensions exist, the extensions list is terminated by the Null TLV, having Type = 0 and Length = 0.

Extensions may occur in any order (see Section 5.3.4 for the exception), but any particular extension type may occur only once in an NHRP packet with the exception of the Vendor Private extension. The vendor-private extension may occur multiple times in a packet in order to allow for extensions which do not share the same vendor ID to be represented. It is RECOMMENDED that a given vendor include no more than one Vendor Private Extension.

An NHS MUST NOT change the order of extensions. That is, the order of extensions placed in an NHRP packet by an NHC (or by an NHS when an NHS sources a packet) MUST be preserved as the packet moves between NHSs. Minimal NHC implementations MUST only recognize, but not necessarily parse, the Vendor Private extension and the End Of Extensions extension. Extensions are only present in a "reply" if they were present in the corresponding "request" with the exception of Vendor Private extensions. The previous statement is not intended to preclude the creation of NHS-only extensions which might be added to and removed from NHRP packets by the same NHS; such extensions MUST not be propagated to NHCs.

The Compulsory bit provides for a means to add to the extension set. If the bit is set, the NHRP message cannot be properly processed by the station responding to the message (e.g., the station that would issue a NHRP Resolution Reply in response to a NHRP Resolution Request) without processing the extension. As a result, the responder MUST return an NHRP Error Indication of type Unrecognized Extension. If the Compulsory bit is clear then the extension can be safely ignored; however, if an ignored extension is in a "request" then it MUST be returned, unchanged, in the corresponding "reply" packet type.

If a transit NHS (one which is not going to generate a "reply") detects an unrecognized extension, it SHALL ignore the extension. If the Compulsory bit is set, the transit NHS MUST NOT cache the information contained in the packet and MUST NOT identify itself as an egress router (in the Forward Record or Reverse Record extensions). Effectively, this means, if a transit NHS encounters an extension which it cannot process and which has the Compulsory bit set then that NHS MUST NOT participate in any way in the protocol exchange other than acting as a forwarding agent.

The NHRP extension Type space is subdivided to encourage use outside the IETF.

0x0000 - 0x0FFF	Reserved for NHRP.
0x1000 - 0x11FF	Allocated to the ATM Forum.
0x1200 - 0x37FF	Reserved for the IETF.
0x3800 - 0x3FFF	Experimental use.

IANA will administer the ranges reserved for the IETF. Values in the ’Experimental use’ range have only local sigificance.

5.3.0 The End Of Extensions

Compulsory = 1
Type = 0
Length = 0

When extensions exist, the extensions list is terminated by the End Of Extensions/Null TLV.

5.3.1 Responder Address Extension

Compulsory = 1
Type = 3
Length = variable

This extension is used to determine the address of the NHRP responder; i.e., the entity that generates the appropriate "reply" packet for a given "request" packet. In the case of an NHRP Resolution Request, the station responding may be different (in the case of cached replies) than the system identified in the Next Hop field of the NHRP Resolution Reply. Further, this extension may aid in detecting loops in the NHRP forwarding path.

This extension uses a single CIE with the extension specific meanings of the fields set as follows:

The Prefix Length fields MUST be set to 0 and ignored.

CIE Code
  5 - Insufficient Resources
    If the responder to an NHRP Resolution Request is an egress point for the target of the address resolution request (i.e., it is one of the stations identified in the list of CIEs in an NHRP Resolution Reply) and the Responder Address extension is included in the NHRP Resolution Request and insufficient resources to setup a cut-through VC exist at the responder then the Code field of the Responder Address Extension is set to 5 in order to tell the client that a VC setup attempt would in all likelihood be rejected; otherwise this field MUST be coded as a zero. NHCs MAY use this field to influence whether they attempt to setup a cut-through to the egress router.

Maximum Transmission Unit
  This field gives the maximum transmission unit preferred by the responder. If this value is 0 then either the default MTU is used or the MTU negotiated via signaling is used if such negotiation is possible for the given NBMA.

Holding Time
  The Holding Time field specifies the number of seconds for which the NBMA information of the responser is considered to be valid. Cached information SHALL be discarded when the holding time expires.

"Client Address" information is actually "Responder Address" information for this extension. Thus, for example, Cli Addr T/L is the responder NBMA address type and length field.

If a "requester" desires this information, the "requester" SHALL include this extension with a value of zero. Note that this implies that no storage is allocated for the Holding Time and Type/Length fields until the "Value" portion of the extension is filled out.

If an NHS is generating a "reply" packet in response to a "request" containing this extension, the NHS SHALL include this extension, containing its protocol address in the "reply". If an NHS has more than one protocol address, it SHALL use the same protocol address consistently in all of the Responder Address, Forward Transit NHS Record, and Reverse Transit NHS Record extensions. The choice of which of several protocol address to include in this extension is a local matter.

If an NHRP Resolution Reply packet being forwarded by an NHS contains a protocol address of that NHS in the Responder Address Extension then that NHS SHALL generate an NHRP Error Indication of type "NHRP Loop Detected" and discard the NHRP Resolution Reply.

If an NHRP Resolution Reply packet is being returned by an intermediate NHS based on cached data, it SHALL place its own address in this extension (differentiating it from the address in the Next Hop field).

5.3.2 NHRP Forward Transit NHS Record Extension

Compulsory = 1
Type = 4
Length = variable

The NHRP Forward Transit NHS record contains a list of transit NHSs through which a "request" has traversed. Each NHS SHALL append to the extension a Forward Transit NHS element (as specified below) containing its Protocol address The extension length field and the ar$chksum fields SHALL be adjusted appropriately.

The responding NHS, as described in Section 5.3.1, SHALL NOT update this extension.

In addition, NHSs that are willing to act as egress routers for packets from the source to the destination SHALL include information about their NBMA Address.

This extension uses a single CIE with the extension specific meanings of the fields set as follows:

The Prefix Length fields MUST be set to 0 and ignored.

CIE Code
5 - Insufficient Resources
If an NHRP Resolution Request contains an NHRP Forward Transit NHS Record Extension and insufficient resources to setup a cut-

through VC exist at the current transit NHS then the CIE Code field for NHRP Forward Transit NHS Record Extension is set to 5 in order to tell the client that a VC setup attempt would in all likelihood be rejected; otherwise this field MUST be coded as a zero. NHCs MAY use this field to influence whether they attempt to setup a cut-through as described in Section 2.2. Note that the NHRP Reverse Transit NHS Record Extension MUST always have this field set to zero.

Maximum Transmission Unit
This field gives the maximum transmission unit preferred by the transit NHS. If this value is 0 then either the default MTU is used or the MTU negotiated via signaling is used if such negotiation is possible for the given NBMA.

Holding Time
The Holding Time field specifies the number of seconds for which the NBMA information of the transit NHS is considered to be valid. Cached information SHALL be discarded when the holding time expires.

"Client Address" information is actually "Forward Transit NHS Address" information for this extension. Thus, for example, Cli Addr T/L is the transit NHS NBMA address type and length field.

If a "requester" wishes to obtain this information, it SHALL include this extension with a length of zero. Note that this implies that no storage is allocated for the Holding Time and Type/Length fields until the "Value" portion of the extension is filled out.

If an NHS has more than one Protocol address, it SHALL use the same Protocol address consistently in all of the Responder Address, Forward NHS Record, and Reverse NHS Record extensions. The choice of which of several Protocol addresses to include in this extension is a local matter.

If a "request" that is being forwarded by an NHS contains the Protocol Address of that NHS in one of the Forward Transit NHS elements then the NHS SHALL generate an NHRP Error Indication of type "NHRP Loop Detected" and discard the "request".

5.3.3 NHRP Reverse Transit NHS Record Extension

Compulsory = 1
Type = 5
Length = variable

The NHRP Reverse Transit NHS record contains a list of transit NHSs through which a "reply" has traversed. Each NHS SHALL append a Reverse Transit NHS element (as specified below) containing its Protocol address to this extension. The extension length field and ar$chksum SHALL be adjusted appropriately.

The responding NHS, as described in Section 5.3.1, SHALL NOT update this extension.

In addition, NHSs that are willing to act as egress routers for packets from the source to the destination SHALL include information about their NBMA Address.

This extension uses a single CIE with the extension specific meanings of the fields set as follows:

The CIE Code and Prefix Length fields MUST be set to 0 and ignored.

Maximum Transmission Unit
  This field gives the maximum transmission unit preferred by the transit NHS. If this value is 0 then either the default MTU is used or the MTU negotiated via signaling is used if such negotiation is possible for the given NBMA.

Holding Time
  The Holding Time field specifies the number of seconds for which the NBMA information of the transit NHS is considered to be valid. Cached information SHALL be discarded when the holding time expires.

"Client Address" information is actually "Reverse Transit NHS Address" information for this extension. Thus, for example, Cli Addr T/L is the transit NHS NBMA address type and length field.

If a "requester" wishes to obtain this information, it SHALL include this extension with a length of zero. Note that this implies that no storage is allocated for the Holding Time and Type/Length fields until the "Value" portion of the extension is filled out.

If an NHS has more than one Protocol address, it SHALL use the same Protocol address consistently in all of the Responder Address, Forward NHS Record, and Reverse NHS Record extensions. The choice of which of several Protocol addresses to include in this extension is a local matter.

If a "reply" that is being forwarded by an NHS contains the Protocol Address of that NHS in one of the Reverse Transit NHS elements then

the NHS SHALL generate an NHRP Error Indication of type "NHRP Loop Detected" and discard the "reply".

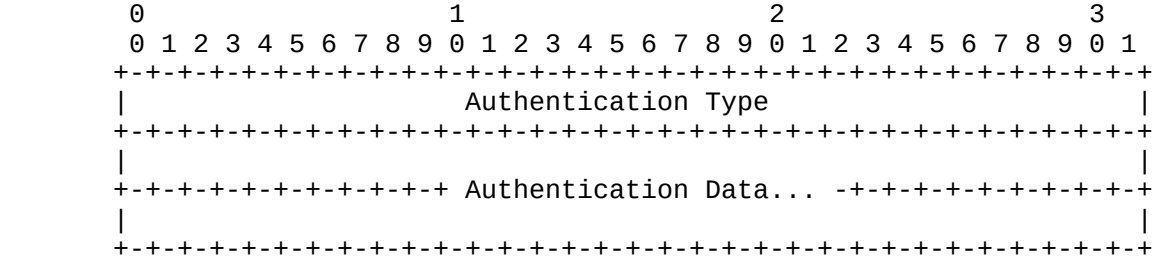
Note that this information may be cached at intermediate NHSs; if so, the cached value SHALL be used when generating a reply.

5.3.4 NHRP Authentication Extension

Compulsory = 1
Type = 7
Length = variable

The NHRP Authentication Extension is carried in NHRP packets to convey authentication information between NHRP speakers. The Authentication Extension may be included in any NHRP "request" or "reply" only.

Except in the case of an NHRP Registration Request/Reply Authentication is done pairwise on an NHRP hop-by-hop basis; i.e., the authentication extension is regenerated at each hop. In the case of an NHRP Registration Request/Reply, the Authentication is checked on an end-to-end basis rather than hop-by-hop. If a received packet fails the authentication test, the station SHALL generate an Error Indication of type "Authentication Failure" and discard the packet. Note that one possible authentication failure is the lack of an Authentication Extension; the presence or absence of the Authentication Extension is a local matter.



The Authentication Type field identifies the authentication method in use. Currently assigned values are:

1 - Cleartext Password
2 - Keyed MD5

All other values are reserved.

The Authentication Data field contains the type-specific

authentication information.

In the case of Cleartext Password Authentication, the Authentication Data consists of a variable length password.

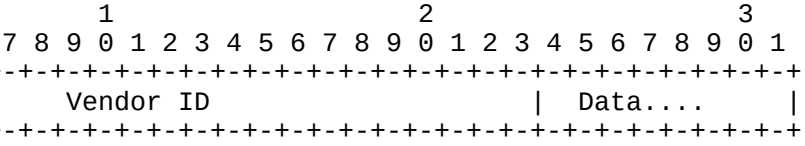
In the case of Keyed MD5 Authentication, the Authentication Data contains the 16 byte MD5 digest of the entire NHRP packet, including the encapsulated protocol's header, with the authentication key appended to the end of the packet. The authentication key is not transmitted with the packet. The MD5 digest covers only the following fields of the NHRP packet: fixed part (with hop count, packet size and checksum being set to zero), mandatory part, Responder Address extension, and authentication extension. Note that when MD5 is used, there is an explicit ordering of the extensions such that: if the Responder Address extension exists then it MUST be the first extension in the packet and the Authentication Extension MUST be the second extension otherwise the Authentication Extension MUST be the first extension in the packet.

Distribution of authentication keys is outside the scope of this document.

5.3.5 NHRP Vendor-Private Extension

Compulsory = 0
Type = 8
Length = variable

The NHRP Vendor-Private Extension is carried in NHRP packets to convey vendor-private information or NHRP extensions between NHRP speakers.



Vendor ID
802 Vendor ID as assigned by the IEEE [6]

Data
The remaining octets after the Vendor ID in the payload are vendor-dependent data.

This extension may be added to any "request" or "reply" packet and it is the only extension that may be included multiple times. If the

receiver does not handle this extension, or does not match the Vendor ID in the extension then the extension may be completely ignored by the receiver. If a Vendor Private Extension is included in a "request" then is must be copied in the corresponding "reply".

6. Protocol Operation

In this section, we discuss certain operational considerations of NHRP.

6.1 Router-to-Router Operation

In practice, the initiating and responding stations may be either hosts or routers. However, there is a possibility under certain conditions that a stable routing loop may occur if NHRP is used between two routers. In particular, attempting to establish an NHRP path across a boundary where information used in route selection is lost may result in a routing loop. Such situations include the loss of BGP path vector information, the interworking of multiple routing protocols with dissimilar metrics (e.g, RIP and OSPF), etc. In such circumstances, NHRP should not be used. This situation can be avoided if there are no "back door" paths between the entry and egress router outside of the NBMA subnetwork. Protocol mechanisms to relax these restrictions are under investigation.

In general it is preferable to use mechanisms, if they exist, in routing protocols to resolve the egress point when the destination lies outside of the NBMA subnetwork, since such mechanisms will be more tightly coupled to the state of the routing system and will probably be less likely to create loops.

6.2 Cache Management Issues

The management of NHRP caches in the source station, the NHS serving the destination, and any intermediate NHSs is dependent on a number of factors.

6.2.1 Caching Requirements

Source Stations

Source stations MUST cache all received NHRP Resolution Replies that they are actively using. They also must cache "incomplete" entries, i.e., those for which a NHRP Resolution Request has been sent but those for which an NHRP Resolution Reply has not been

received. This is necessary in order to preserve the Request ID for retries, and provides the state necessary to avoid triggering NHRP Resolution Requests for every data packet sent to the destination.

Source stations MUST purge expired information from their caches. Source stations MUST purge the appropriate cached information upon receipt of an NHRP Purge Request packet.

When a station has a co-resident NHC and NHS, the co-resident NHS may reply to NHRP Resolution Requests from the co-resident NHC with information which the station cached as a result of the co-resident NHC making its own NHRP Resolution Requests as long as the co-resident NHS follows the rules for Transit NHSs as seen below.

Serving NHSs

The NHS serving the destination (the one which responds authoritatively to NHRP Resolution Requests) SHOULD cache protocol address information from all NHRP Resolution Requests to which it has responded if the information in the NHRP Resolution Reply has the possibility of changing during its lifetime (so that an NHRP Purge Request packet can be issued). The internetworking to NBMA binding information provided by the source station in the NHRP Resolution Request may also be cached if and only if the "S" bit is set, the NHRP Resolution Request has included a CIE with the Holding Time field set greater than zero (this is the valid Holding Time for the source binding), and only for non-authoritative use for a period not to exceed the Holding Time.

Transit NHSs

A Transit NHS (lying along the NHRP path between the source station and the responding NHS) may cache source binding information contained in NHRP Resolution Request packets that it forwards if and only if the "S" bit is set, the NHRP Resolution Request has included a CIE with the Holding Time field set greater than zero (this is the valid Holding Time for the source binding), and only for non-authoritative use for a period not to exceed the Holding Time.

A Transit NHS may cache destination information contained in NHRP Resolution Reply CIE if only if the D bit is set and then only for non-authoritative use for a period not to exceed the Holding Time value contained in the CIE. A Transit NHS MUST NOT cache source binding information contained in an NHRP Resolution Reply.

Further, a transit NHS MUST discard any cached information when the

prescribed time has expired. It may return cached information in response to non-authoritative NHRP Resolution Requests only.

6.2.2 Dynamics of Cached Information

NBMA-Connected Destinations

NHRP's most basic function is that of simple NBMA address resolution of stations directly attached to the NBMA subnetwork. These mappings are typically very static, and appropriately chosen holding times will minimize problems in the event that the NBMA address of a station must be changed. Stale information will cause a loss of connectivity, which may be used to trigger an authoritative NHRP Resolution Request and bypass the old data. In the worst case, connectivity will fail until the cache entry times out.

This applies equally to information marked in NHRP Resolution Replies as being "stable" (via the "D" bit).

Destinations Off of the NBMA Subnetwork

If the source of an NHRP Resolution Request is a host and the destination is not directly attached to the NBMA subnetwork, and the route to that destination is not considered to be "stable," the destination mapping may be very dynamic (except in the case of a subnetwork where each destination is only singly homed to the NBMA subnetwork). As such the cached information may very likely become stale. The consequence of stale information in this case will be a suboptimal path (unless the internetwork has partitioned or some other routing failure has occurred).

6.3 Use of the Prefix Length field of a CIE

A certain amount of care needs to be taken when using the Prefix Length field of a CIE, in particular with regard to the prefix length advertised (and thus the size of the equivalence class specified by it). Assuming that the routers on the NBMA subnetwork are exchanging routing information, it should not be possible for an NHS to create a black hole by advertising too large of a set of destinations, but suboptimal routing (e.g., extra internetwork layer hops through the NBMA) can result. To avoid this situation an NHS that wants to send the Prefix Length MUST obey the following rule:

The NHS examines the Network Layer Reachability Information (NLRI) associated with the route that the NHS would use to forward towards the destination (as specified by the Destination internetwork layer

address in the NHRP Resolution Request), and extracts from this NLRI the shortest address prefix such that: (a) the Destination internetwork layer address (from the NHRP Resolution Request) is covered by the prefix, (b) the NHS does not have any routes with NLRI which form a subset of what is covered by the prefix. The prefix may then be used in the CIE.

The Prefix Length field of the CIE should be used with restraint, in order to avoid NHRP stations choosing suboptimal transit paths when overlapping prefixes are available. This document specifies the use of the prefix length only when all the destinations covered by the prefix are "stable". That is, either:

(a) All destinations covered by the prefix are on the NBMA network, or

(b) All destinations covered by the prefix are directly attached to the NHRP responding station.

Use of the Prefix Length field of the CIE in other circumstances is outside the scope of this document.

6.4 Domino Effect

One could easily imagine a situation where a router, acting as an ingress station to the NBMA subnetwork, receives a data packet, such that this packet triggers an NHRP Resolution Request. If the router forwards this data packet without waiting for an NHRP transit path to be established, then when the next router along the path receives the packet, the next router may do exactly the same - originate its own NHRP Resolution Request (as well as forward the packet). In fact such a data packet may trigger NHRP Resolution Request generation at every router along the path through an NBMA subnetwork. We refer to this phenomena as the NHRP "domino" effect.

The NHRP domino effect is clearly undesirable. At best it may result in excessive NHRP traffic. At worst it may result in an excessive number of virtual circuits being established unnecessarily. Therefore, it is important to take certain measures to avoid or suppress this behavior. NHRP implementations for NHSs MUST provide a mechanism to address this problem. One possible strategy to address this problem would be to configure a router in such a way that NHRP Resolution Request generation by the router would be driven only by the traffic the router receives over its non-NBMA interfaces (interfaces that are not attached to an NBMA subnetwork). Traffic received by the router over its NBMA-attached interfaces would not trigger NHRP Resolution Requests. Such a router avoids the NHRP domino effect through administrative means.

7. NHRP over Legacy BMA Networks

There would appear to be no significant impediment to running NHRP over legacy broadcast subnetworks. There may be issues around running NHRP across multiple subnetworks. Running NHRP on broadcast media has some interesting possibilities; especially when setting up a cut-through for inter-ELAN inter-LIS/LAG traffic when one or both end stations are legacy attached. This use for NHRP requires further research.

8. Security Considerations

As in any resolution protocol, there are a number of potential security attacks possible. Plausible examples include denial-of-service attacks, and masquerade attacks using register and purge packets. The use of authentication on all packets is recommended to avoid such attacks.

The authentication schemes described in this document are intended to allow the receiver of a packet to validate the identity of the sender; they do not provide privacy or protection against replay attacks.

Detailed security analysis of this protocol is for further study.

9. Discussion

The result of an NHRP Resolution Request depends on how routing is configured among the NHSs of an NBMA subnetwork. If the destination station is directly connected to the NBMA subnetwork and the the routed path to it lies entirely within the NBMA subnetwork, the NHRP Resolution Replies always return the NBMA address of the destination station itself rather than the NBMA address of some egress router. On the other hand, if the routed path exits the NBMA subnetwork, NHRP will be unable to resolve the NBMA address of the destination, but rather will return the address of the egress router. For destinations outside the NBMA subnetwork, egress routers and routers in the other subnetworks should exchange routing information so that the optimal egress router may be found.

In addition to NHSs, an NBMA station could also be associated with one or more regular routers that could act as "connectionless servers" for the station. The station could then choose to resolve the NBMA next hop or just send the packets to one of its connectionless servers. The latter option may be desirable if communication with the destination is short-lived and/or doesn’t

require much network resources. The connectionless servers could, of course, be physically integrated in the NHSs by augmenting them with internetwork layer switching functionality.

References

[1] NBMA Address Resolution Protocol (NARP), Juha Heinanen and Ramesh Govindan, RFC1735.

[2] Address Resolution Protocol, David C. Plummer, RFC 826.

[3] Classical IP and ARP over ATM, Mark Laubach, RFC 1577.

[4] Transmission of IP datagrams over the SMDS service, J. Lawrence and D. Piscitello, RFC 1209.

[5] Protocol Identification in the Network Layer, ISO/IEC TR 9577:1990.

[6] Assigned Numbers, J. Reynolds and J. Postel, RFC 1700.

[7] Multiprotocol Encapsulation over ATM Adaptation Layer 5, J. Heinanen, RFC1483.

[8] Multiprotocol Interconnect on X.25 and ISDN in the Packet Mode, A. Malis, D. Robinson, and R. Ullmann, RFC1356.

[9] Multiprotocol Interconnect over Frame Relay, T. Bradley, C. Brown, and A. Malis, RFC1490.

[10] "Local/Remote" Forwarding Decision in Switched Data Link Subnetworks, Yakov Rekhter, Dilip Kandlur, RFC1937.

[11] Support for Multicast over UNI 3.0/3.1 based ATM Networks, G. Armitage, Work In Progress.

[12] Server Cache Synchronization Protocol (SCSP) - NBMA, J. Luciani, G. Armitage, J. Halpern, Work In Progress.

[13] NHRP for Destinations off the NBMA Subnetwork, Y. Rekhter, Work In Progress.

[14] Classical IP to NHRP Transition, J. Luciani, et al., Work In Progress.

Acknowledgments

We would like to thank (in no particular order) Juha Heinenan of Telecom Finland and Ramesh Govidan of ISI for their work on NBMA ARP and the original NHRP draft, which served as the basis for this work. Russell Gardo of IBM, John Burnett of Adaptive, Dennis Ferguson of ANS, Andre Fredette of Bay Networks, Joel Halpern of Newbridge, Paul Francis of NTT, Tony Li, Bryan Gleeson, and Yakov Rekhter of cisco, and Grenville Armitage of Bellcore should also be acknowledged for comments and suggestions that improved this work substantially. We would also like to thank the members of the ION working group of the IETF, whose review and discussion of this document have been invaluable.

Authors' Addresses

James V. Luciani
Bay Networks
3 Federal Street
Mail Stop: BL3-04
Billerica, MA 01821
Phone: +1 508 916 4734
Email: luciani@baynetworks.com

Dave Katz
cisco Systems
170 W. Tasman Dr.
San Jose, CA 95134 USA
Phone: +1 408 526 8284
Email: dkatz@cisco.com

David Piscitello
Core Competence
1620 Tuckerstown Road
Dresher, PA 19025 USA
Phone: +1 215 830 0692
Email: dave@corecom.com

Bruce Cole
Juniper Networks
3260 Jay St.
Santa Clara, CA 95054
Phone: +1 408 327 1900
Email: bcole@jnx.com

Appendix I. State Machine View of MPOA Component Behavior [Informative]

A state machine view of the MPOA component behavior is given in this Section. The state machines shown are intended to give an example of a compliant implementation, but do not represent the complete specification.

I.1 Conventions

Vertical lines represent states, the names of which are labeled above each vertical line. Transitions are horizontal lines. Events are labeled above each transition. Actions are labeled below each transition. The Idle state is shown as a dashed line.

I.2 Ingress MPC Control State Machine

Each instance of the ingress MPC State Machine is defined in the context of an ingress cache entry, namely the <MPS Control ATM Address, Destination Internetwork Layer Address> tuple.

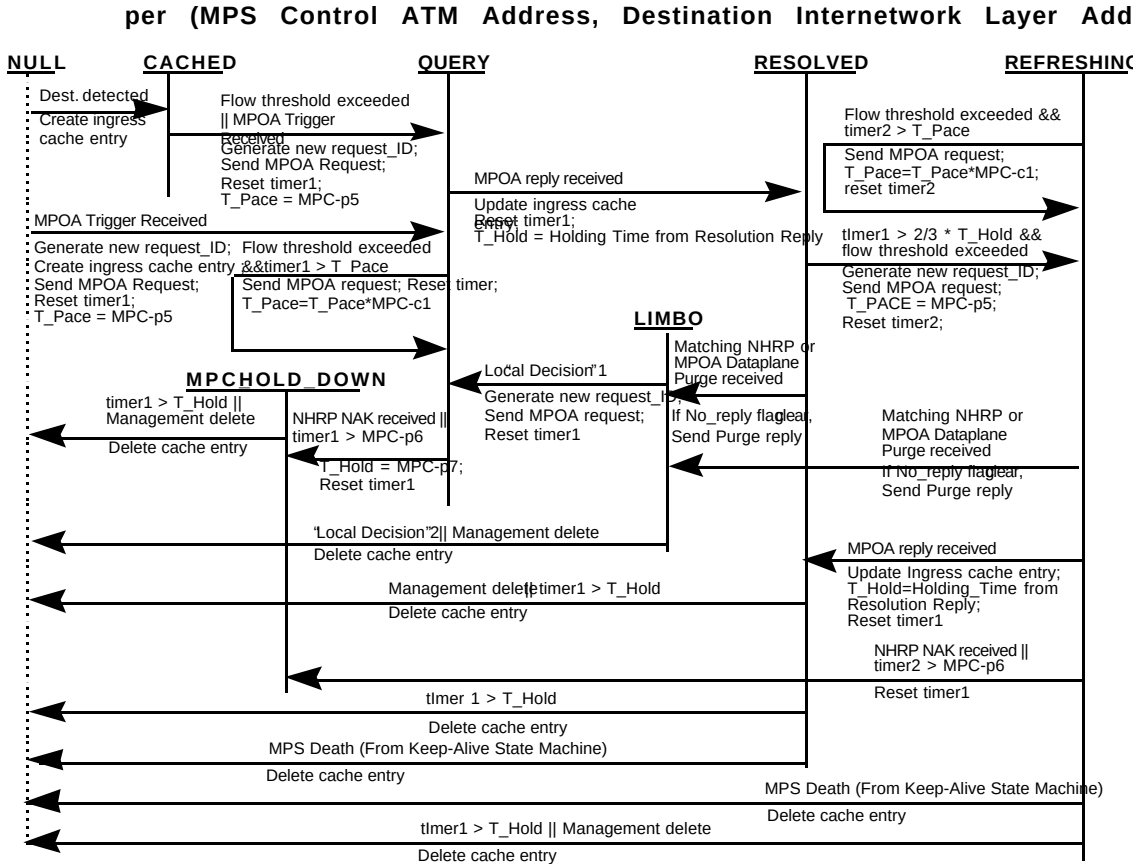


Figure 23 Ingress MPC Control State Machine

Notes on Figure 23:

1. timer1 and timer2 are second timers. Two are required to handle hold times and retry times in the REFRESHING state. Retries are paced until MPC-p5 is reached, in which case the cache entry is deleted when the cache hold time is reached. Alternatively, retries could be paced until the hold time is reached.

2. MPOA trigger stimulates an MPOA request from both NULL and CACHED states.
3. Local Decision 1 and 2 refer to text in Section 4.4.4 on Cache Management

I.3 Ingress MPS Control State Machine

Each instance of the ingress MPS state machine is defined by the <Ingress MPC control ATM Address, Request ID> tuple obtained from the MPOA Resolution Request. The purge description applies to a single cache entry for a single ingress MPC and assumes a higher layer fan-out process.

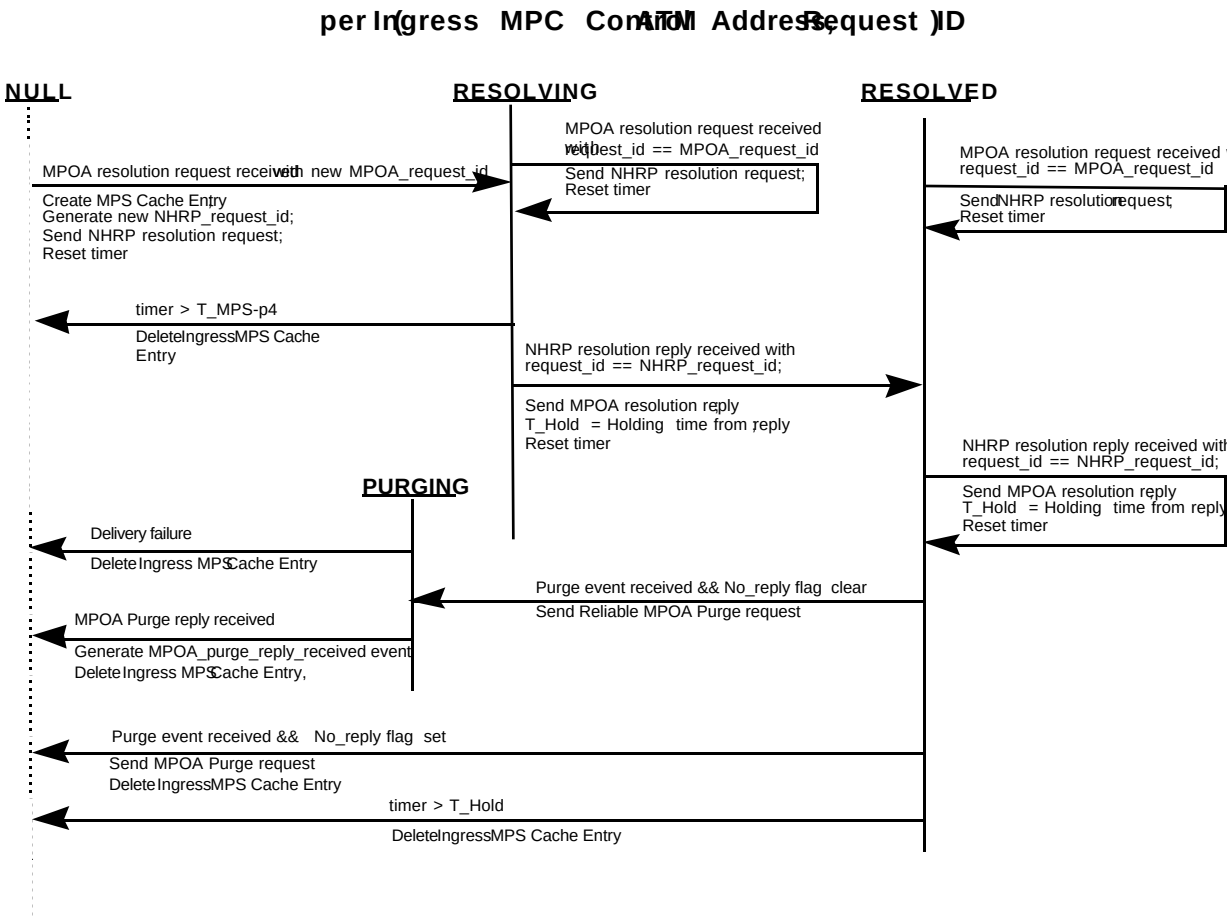


Figure 24 Ingress MPS Control State Machine

Notes on Figure 24:

1. timer is a second timer
2. Purge events are generated by a process outside the scope of this state machine. This process takes a received NHRP purge request message potentially containing summarized information, and generates events corresponding to instances of this state machine. This process stores state necessary for eventual generation of a NHRP purge reply message.
3. The MPOA_purge_reply_received event is sent from this state machine to the above process. This process uses this event data to determine if and when to generate an NHRP purge reply.
4. The Delivery failure event is generated when reliable delivery fails.

I.4 Egress MPS Control State Machine

per (Source ATM Address, Destination Internetwork Layer Addr

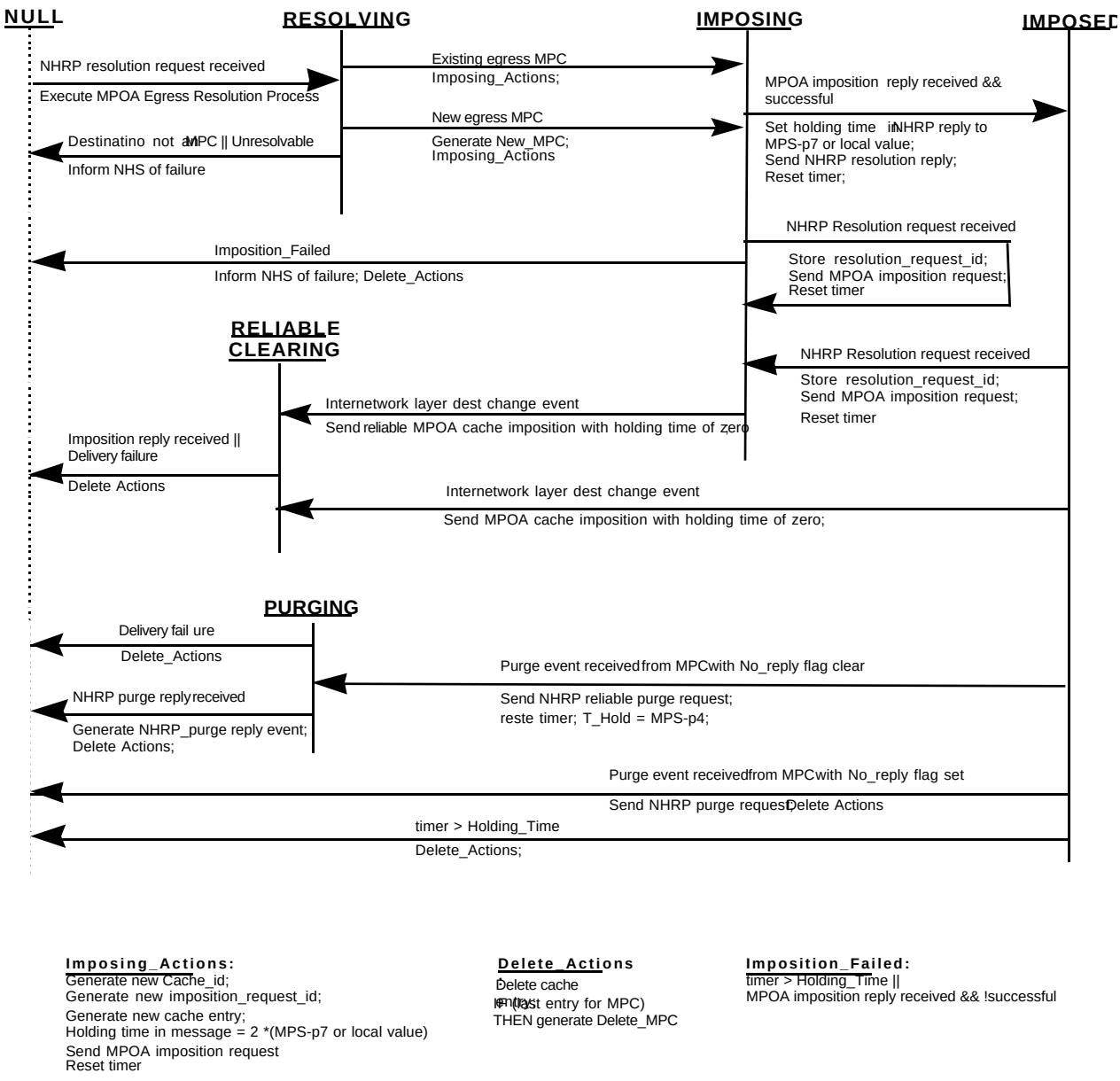


Figure 25 Egress MPS Control State Machine

Notes on Figure 25:

1. timer is a second timer.
2. The purge_event is generated by a process outside the context of this state machine, since the MPOA purge request from the MPS may contain summary information. If the N bit is clear(indicating that a reply is needed) , a purge_reply _event is sent to this process when a corresponding NHRP reply is received.
3. The NHS is informed of an imposition failure, so that it may terminate the shortcut if it desires.
4. Internetwork layer dest change event is generated by a process out of the context of this state machine.

5. Holding_Time is derived from either MPS-p7 or 'local information'. Refer to Section 4.1.1.1.

I.5 Egress MPC Control State Machine

(per Cache_ID, MPS Control ATM Address)

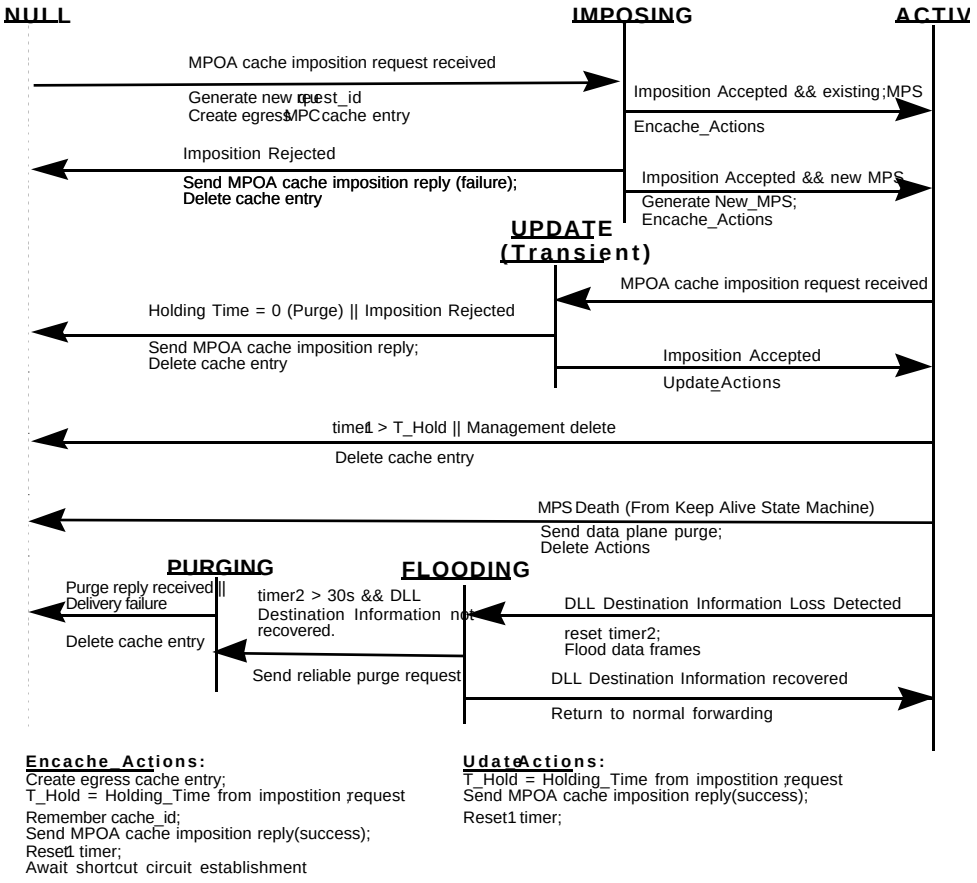


Figure 26 Egress MPC Control State Machine

Notes on Figure 26:

1. Purge event is generated by a process outside the context of this state machine
2. The change in DLL destination state is generated outside the context of this state machine. This event is generated when there is a change in one ore more the following LEC variables: C6, C8, C27, C30

I.6 Reliable Delivery State Machines

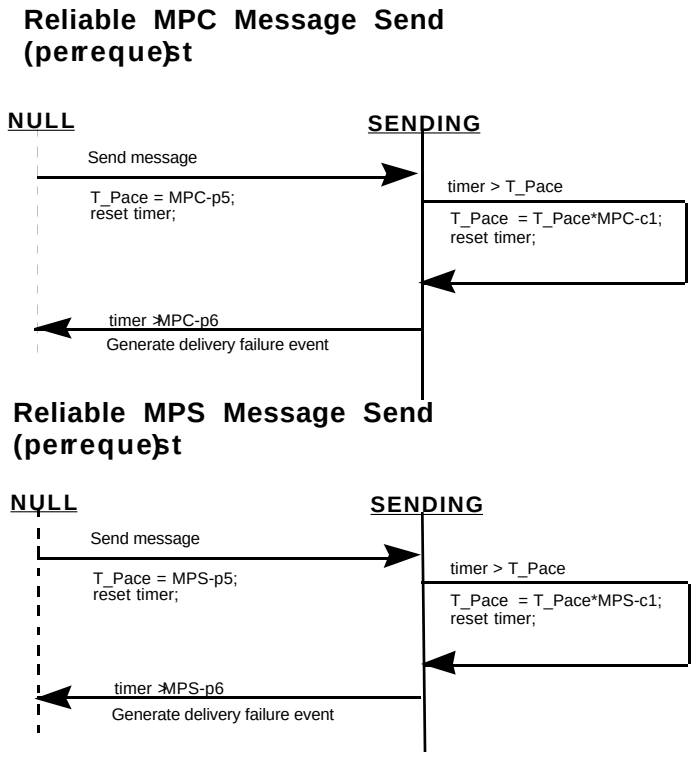


Figure 27 Reliable Delivery State Machines

Notes on Figure 27:

1. timer is a second timer
2. Purge event is generated by a process outside the context of this state machine
3. The change in destination state is generated outside the context of this state machine. This event is generated when there is a change in one ore more the following LEC variables: C6, C8, C27, C30

I.7 Egress MPC and MPS Keep-Alive State Machines

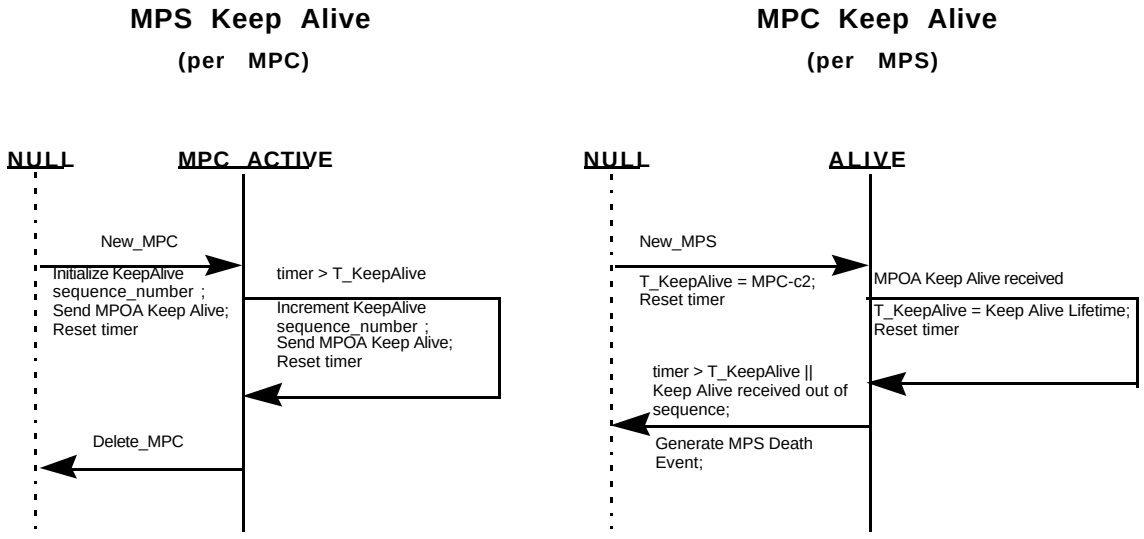


Figure 28 Egress MPC and MPS Keep-Alive State Machines

Notes on Figure 28:

1. timer is a second timer.
2. Purge text described above is from Section 4.5.

Appendix II. Examples of MPOA Control and Data Flows [Informative]

II.1 Scenarios

A simple MPOA network configuration is shown in Figure 29. The MPOA network consists of two ELANs: ELAN-1 and ELAN-2. Each ELAN contains one or more edge devices and MPOA hosts. Each edge device supports one or more LAN hosts.

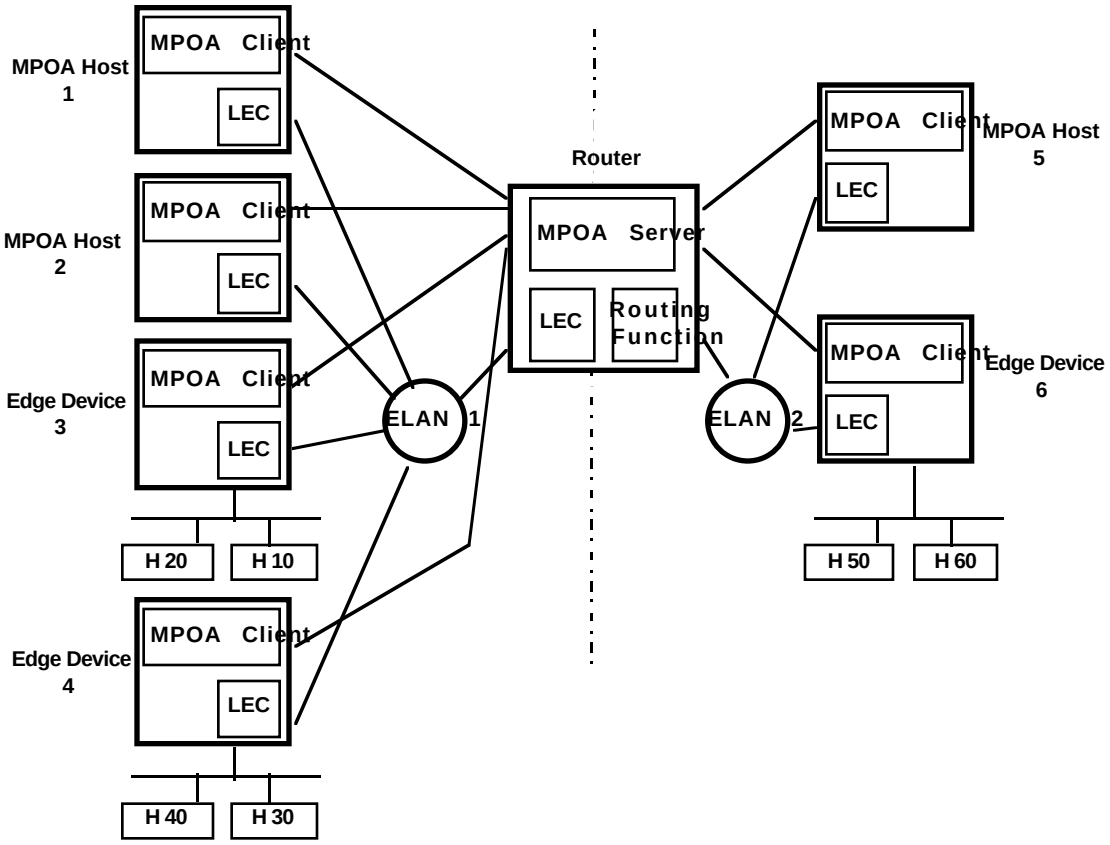


Figure 29 Example Network Configuration

To describe each flow, a source-destination pair (an MPOA host and/or a LAN host) is chosen from Figure 29. The source and destination are chosen within the same ELAN or in different ELANs and the flows are grouped as the intra-ELAN and inter-ELAN flows.

II.1.1 Intra-ELAN Scenarios

Intra-ELAN flows originate from an MPOA host or a LAN host behind an edge device, and flow to an MPOA host or a LAN Host in the same ELAN. These flows use LANE for address resolution and data transfer. The matrix shown in Table 7 illustrates all source-destination pairs with the matrix entry representing the scenario-index. Note that the source and destination are different hosts. The trivial scenario of a LAN-LAN flow on the same edge device is not covered.

Table 7 Intra-ELAN Scenarios

	To MPOA Host	To LAN host

From MPOA Host	(A)	(B)
From LAN host	(C)	(D)

II.1.2 Inter-ELAN Scenarios

The flows listed in Table 8 are between the source-destination pairs for which the source and destination are in different ELANs. These flows may use a default path for short-lived flows or a shortcut for long-lived flows. The default path uses the LANE and router capabilities. The shortcut path uses LANE plus NHRP for address resolution and a shortcut for data transfer.

Table 8 Inter-ELAN Scenarios

	To MPOA Host	**To LAN Host**
From MPOA Host	(E)	(F)
From LAN Host	(G)	(H)

II.2 Flows

II.2.1 Intra-ELAN

Intra-ELAN flows are illustrated in Figure 30.

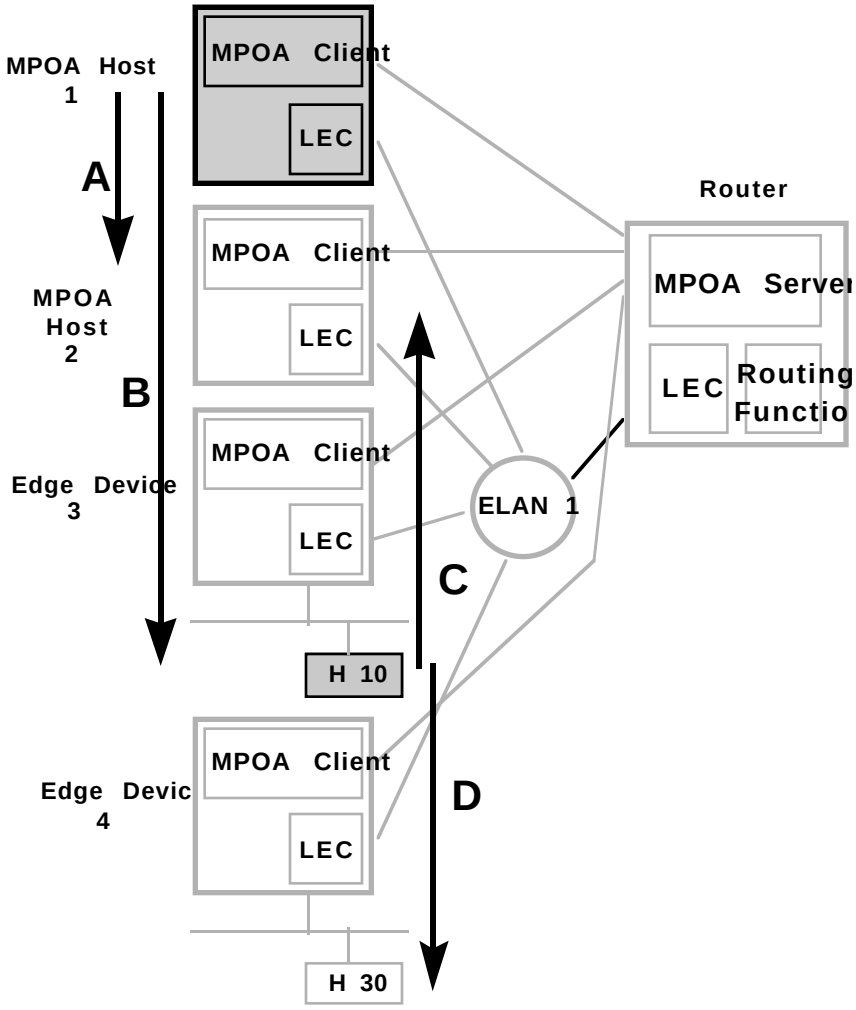


Figure 30 Intra-ELAN Flows

II.2.1.1 From MPOA Host

II.2.1.1.1 Scenario (A): MPOA Host 1 to MPOA Host 2

Figure 31 shows the data path for data originating from MPOA Host 1 and destined to MPOA Host 2 within the same ELAN. LANE is used for such a flow and a Data Direct VCC will carry the LANE frames.

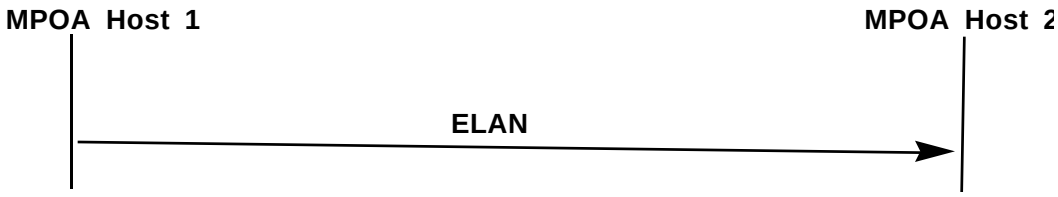


Figure 31 MPOA Host to MPOA Host Data Flow

II.2.1.1.2 Scenario (B): MPOA Host 1 to LAN Host H 10

Figure 32 shows the data path for data originating from MPOA Host 1 and destined to LAN Host H 10 within the same ELAN. LANE is used for such a flow and a Data Direct VCC between MPOA Host 1 and Edge Device 3 will carry the LANE frames.

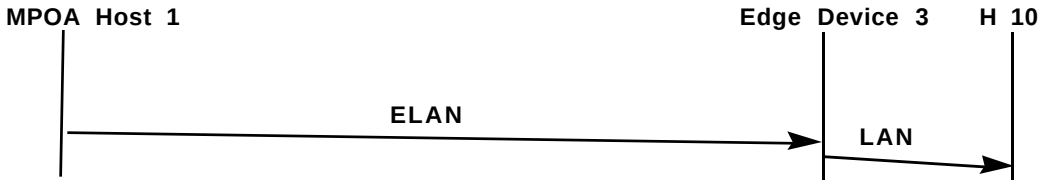


Figure 32 MPOA Host to LAN Host Data Flow

II.2.1.2 From LAN Host

II.2.1.2.1 Scenario (C): LAN Host H 10 to MPOA Host 2

Figure 33 shows the data path for data originating from LAN Host H 10 and destined to MPOA Host 2 within the same ELAN. LANE is used for such a flow and a Data Direct VCC between Edge Device 3 and MPOA Host 2 will carry the LANE frames.

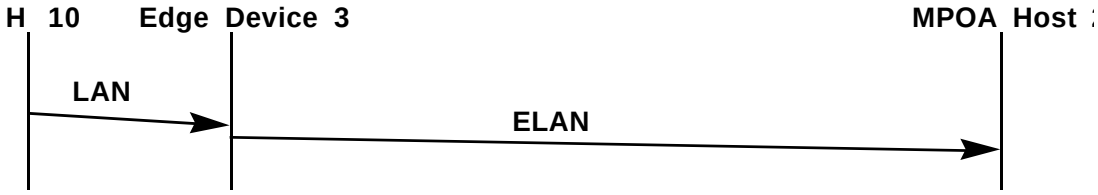


Figure 33 LAN Host to MPOA Host Data Flow

II.2.1.2.2 Scenario (D): LAN Host H 10 to LAN Host H 30

Figure 34 shows the data path for data originating from LAN Host H 10 and destined to LAN Host H 30 within the same ELAN. LANE is used for such a flow and a Data Direct VCC between Edge Device 3 and Edge Device 4 will carry the LANE frames.

Figure 34 LAN Host to LAN Host Data Flow

II.2.2 Inter-ELAN

Inter-ELAN flows are illustrated in Figure 35. For the sake of simplicity, the flows shown in this section only involve a single router/MPS; however, in practice, there can be an arbitrary number of routers on the default routed path between the ingress MPS and egress MPS. The ingress MPS actions (MPOA Resolution Request/Reply) and egress MPS actions (MPOA Cache Imposition Request/Reply) are independent, and in a more complex scenario where multiple routers exist on the default path, the ingress and egress MPSs communicate via NHRP.

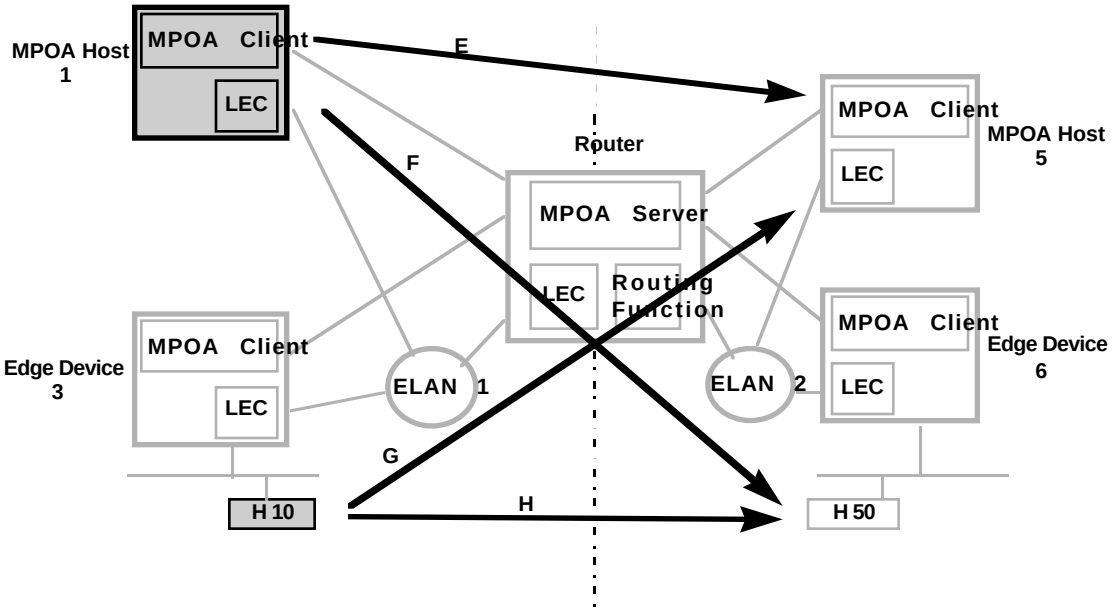


Figure 35 Inter-ELAN Flows

II.2.2.1 From MPOA Host

II.2.2.1.1 Scenario (E): MPOA Host 1 to MPOA Host 5

Figure 36 shows the default and shortcut data paths for data originating from MPOA Host 1 and destined to MPOA Host 5 within a different ELAN.

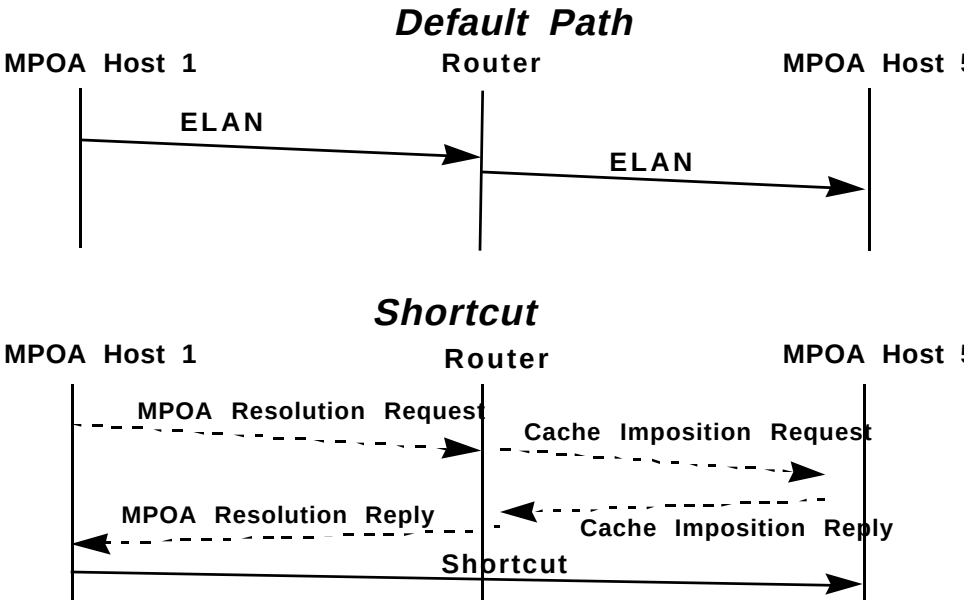


Figure 36 MPOA Host to MPOA Host

Default Path:

MPOA Host 1 sends the packet in a LANE frame to the router via a Data Direct VCC. The router forwards the packet in a LANE frame to MPOA Host 5 via another Data Direct VCC.

Shortcut:

If MPOA Host 1 detects a flow to the internetwork layer address of MPOA Host 5, it sends an MPOA Resolution Request to the MPS to get the corresponding ATM address. The router sends an MPOA Cache Imposition Request to MPOA Host 5 to provide the egress cache entry. MPOA Host 5 sends an MPOA Cache Imposition Reply to the MPS indicating that it can accept the shortcut. The router sends an MPOA Resolution Reply back to MPOA Host 1 with the ATM address of MPOA Host 5. MPOA Host 1 may then update its ingress cache and establish a shortcut to MPOA Host 5.

For subsequent data destined to MPOA Host 5, MPOA Host 1 encapsulates the internetwork layer protocol packet with the appropriate encapsulation for the shortcut. The packets are then sent to MPOA Host 5 using the VCC specified in the ingress cache entry.

II.2.2.1.2 Scenario (F): MPOA Host 1 to LAN Host H 50

Figure 37 shows the default and shortcut data paths for data originating from MPOA Host 1 and destined to LAN Host H 50 within a different ELAN.

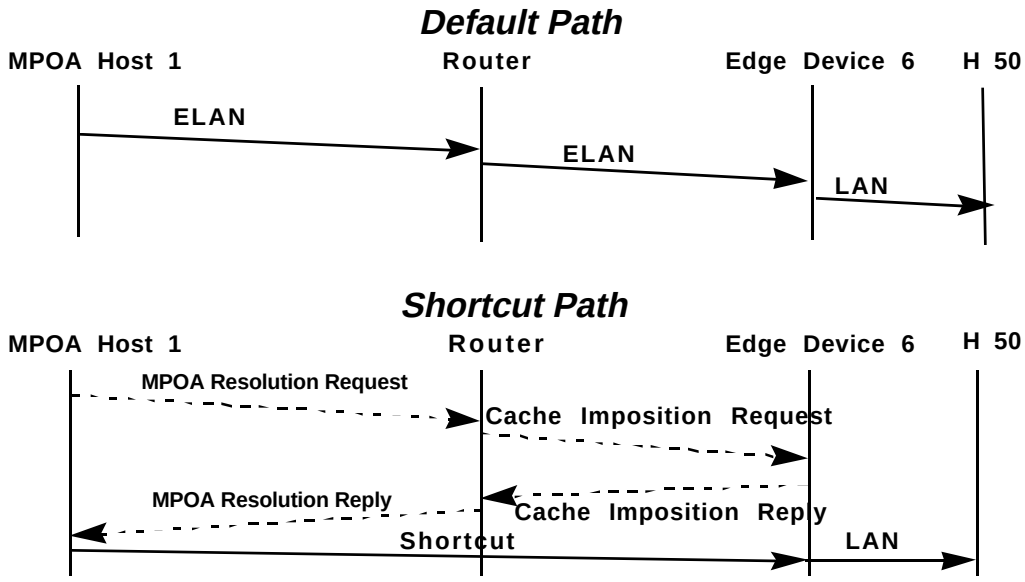


Figure 37 MPOA Host to LAN Host

Default Path:

MPOA Host 1 sends the packet in a LANE frame to the router via a Data Direct VCC. The router forwards the packet in a LANE frame to Edge Device 6 via another Data Direct VCC. Edge Device 6 sends the MAC frame to the LAN Host 50.

Shortcut:

If MPOA Host 1 detects a flow to the internetwork layer address of LAN Host H 50, it sends an MPOA Resolution Request to the MPS to get the corresponding ATM address. The router sends an MPOA Cache Imposition Request to Edge Device 6 to provide the egress cache entry. Edge Device 6 sends an MPOA Cache Imposition Reply to the MPS indicating that it can accept the shortcut. The router sends an MPOA Resolution Reply to MPOA Host 1 with the ATM address of Edge Device 6. MPOA Host 1 may then update its ingress cache and establish a shortcut to Edge Device 6.

For subsequent data destined to LAN Host H 50, MPOA Host 1 encapsulates the internetwork layer protocol packet with the appropriate encapsulation for the shortcut. The packets are then sent to Edge Device 6 using the VCC specified in the cache entry. Edge Device 6 receives the encapsulated packets, makes the MAC frames and sends them to LAN Host 50.

II.2.2.2 From LAN Host

II.2.2.2.1 Scenario (G): LAN Host H 10 to MPOA Host 5

Figure 38 shows the default and shortcut data paths for data originating from LAN Host H 10 and destined to MPOA Host 5 within a different ELAN.

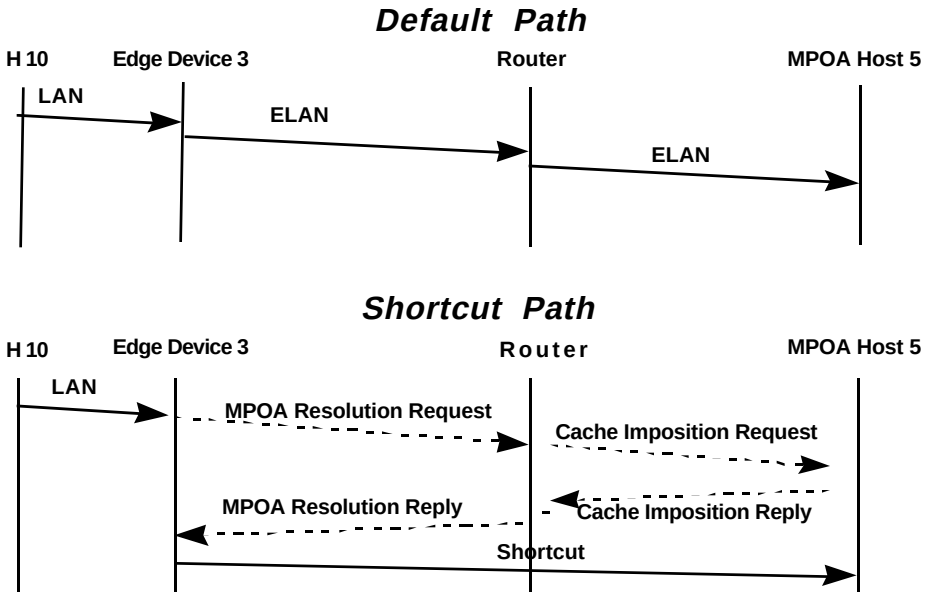


Figure 38 LAN Host to MPOA Host

Default Path:

LAN Host 10 sends the MAC frame to Edge Device 3. Edge Device 3sends the packet in a LANE frame to the router via a Data Direct VCC. The router forwards the packet in a LANE frame to MPOA Host 5 via another Data Direct VCC.

Shortcut:

LAN Host 10 sends the MAC frame to Edge Device 3. If Edge Device 3 detects a flow to the internetwork layer address of MPOA Host 5, it sends an MPOA Resolution Request to the MPS to get the corresponding ATM address. The router sends an MPOA Cache Imposition Request to MPOA Host 5 to provide the egress cache entry. MPOA Host 5 sends an MPOA Cache Imposition Reply to the MPS indicating that it can accept the shortcut. The router sends an MPOA Resolution Reply to Edge Device 3 with the ATM address of MPOA Host 5. Edge Device 3 may then update its ingress cache and establish a shortcut to MPOA Host 5.

For subsequent data destined to MPOA Host 5, Edge Device 3 encapsulates the internetwork layer protocol packet with the appropriate encapsulation for the shortcut. The packets are then sent to MPOA Host 5 using the VCC specified in the cache entry.

II.2.2.2.2 Scenario (H): LAN Host H 10 to LAN Host H 50

Figure 39 shows the default and shortcut data path for data originating from LAN Host H 10 and destined to LAN Host H 50 within a different ELAN.

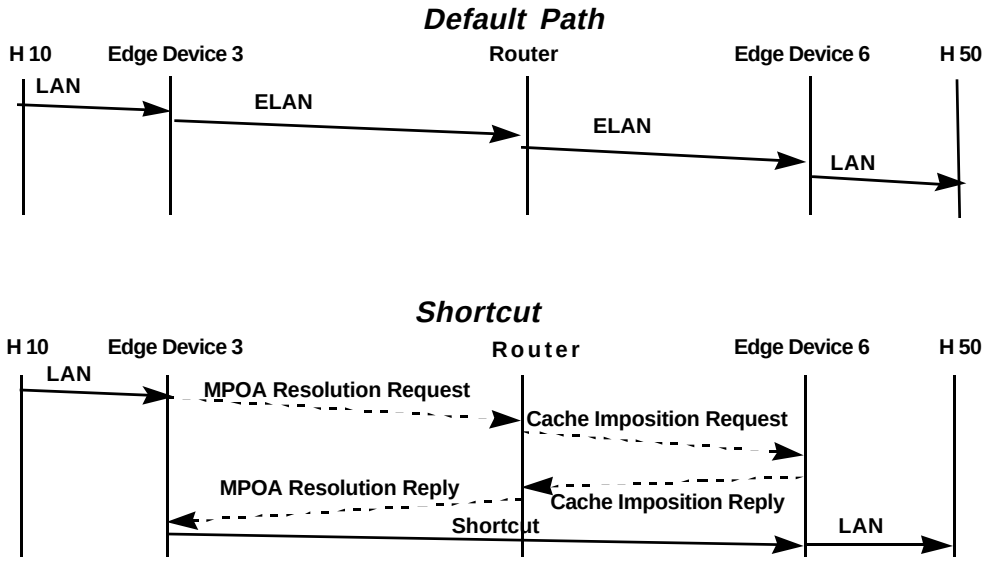


Figure 39 LAN Host to LAN Host

Default Path:

LAN Host 10 sends the MAC frame to Edge Device 3. Edge Device 3 sends the packet in a LANE frame to the router via a Data Direct VCC. The router forwards the packet in a LANE frame to Edge Device 6 via another Data Direct VCC. Edge Device 6 sends the MAC frame to the LAN Host 50.

Shortcut:

LAN Host 10 sends the MAC frame to Edge Device 3. If Edge Device 3 detects a flow to the internetwork layer address of LAN Host H 50, it sends an MPOA Resolution Request to the MPS to get the corresponding ATM address. The router sends an MPOA Cache Imposition Request to Edge Device 6 to provide the egress cache entry. Edge Device 6 sends an MPOA Cache Imposition Reply to the MPS indicating that it can accept the shortcut. The router sends an MPOA Resolution Reply to Edge Device 3 with the ATM address of Edge Device 6. Edge Device 3 may then update its ingress cache and establish a shortcut to MPOA Host 5.

For subsequent data destined to LAN Host H 50, Edge Device 3 encapsulates the internetwork layer protocol packet with the appropriate encapsulation for the shortcut. The packets are then sent to Edge Device 6 using the VCC specified in the cache entry. Edge Device 6 receives the encapsulated packets, makes the MAC frames and sends them to LAN Host 50.

Appendix III. Related Work [Informative]

The rapid and wide acceptance of ATM has stimulated enormous activity in the communications industry to standardize ATM interfaces. One of the principal objectives of this activity is to enable protocols and applications at the internetwork layer and above to operate effectively over an ATM transport network. Enabling existing protocols and applications to operate over ATM is generally viewed as one of the final, necessary steps to allow the benefits of ATM to be brought gradually into existing networks. Several industry projects address different pieces of the internetwork layer problem, including the LAN Emulation over ATM (LANE) specification under the auspices of the ATM Forum, Classical IP and ARP over ATM defined in RFC 1577 (Classical IP), the Next Hop Resolution Protocol (NHRP), the Multicast Address Resolution Protocol (MARS), RFC 1483 and other projects under the auspices of the Internet Engineering Task Force (IETF). The APPN Implementers Workshop (AIW) has addressed extensions for the High Performance Routing (HPR) protocol for ATM networks in [AIW]. These projects have all taken a somewhat similar technical approach. The existing protocol stack is either left unchanged (e.g., LANE), or is modified in only minor ways (e.g., Classical IP). These projects have additionally required that any changes or additions can be made only to protocol stacks in systems that have a direct ATM interface, that no changes can be made to the protocol stacks on "LAN systems" (i.e., systems attached to existing subnetworks), and that ATM-attached systems and LAN systems must be fully interoperable.

III.1 LANE

The LANE specification defines a method for an ATM network to emulate an Ethernet or Token-Ring LAN. Protocols, such as IP's Address Resolution Protocol (ARP), that are dependent on the availability of a broadcast function, are supported by LANE over ATM which, due to its connection-oriented nature, is inherently non-broadcast. A host on a LAN that wishes to send data to another host on that LAN using an internetwork layer protocol must determine the MAC (medium access control) address of the destination host prior to data transfer. Protocols such as ARP use broadcast to resolve internetwork layer addresses to MAC Addresses by querying all end-stations on the LAN. On an Emulated LAN (ELAN), LANE supports this broadcast with a Broadcast/Unknown Server (BUS). However, to effect the actual data transfer over an ATM network, a further mapping from MAC Address to ATM address is necessary. Another server, the LANE server (LES), provides this mapping. The originating host then transfers the data by setting up an ATM VCC to the target ATM address.

III.2 Classical IP

The MAC-to-ATM address resolution provided by LANE, while allowing Internetwork and higher-layer protocols to operate as they do on an Ethernet or token ring LAN, involves two levels of resolution prior to data transfer. An internetwork layer address must first be resolved to a MAC Address, and then the MAC Address is mapped to an ATM address. RFC 1577 is the definition of an enhanced IP ARP procedure that resolves internetwork layer addresses directly to ATM addresses. A server, known as the ATMARP server, responds to queries from hosts for internetwork layer addresses with an ATM address. By reducing one step in the process of setting up an ATM connection for data transfer, RFC 1577 helps to minimize broadcast traffic in the subnet. The ATMARP server provides this service to all IP end-stations that are directly attached to the ATM network in a Logical IP Subnet (or LIS, the scope of which is defined in RFC 1577). RFC 1577 applies only to IP, while LANE supports all internetwork layer protocol.

III.3 MARS

Rounding out the suite of protocols for ATM internetwork layer is the IETF's MARS (Multicast Address Resolution Server) specification. MARS is used to resolve internetwork layer multicast and broadcast addresses to either a list of ATM addresses, or to the ATM address of a Multicast Server (MCS) that is responsible for distributing the data to the appropriate end-stations. A MARS serves end-stations in a MARS *cluster*, which is currently equivalent to a LIS. Further study is required before the scope of a cluster is extended beyond a LIS.

III.4 RFC 1483

RFC 1483, Multiprotocol Encapsulation over AAL5, describes encapsulation mechanisms that higher layer protocols can use for transport using ATM Adaptation Layer 5. Data on ATM VCCs established using any of the above methods may be encapsulated using the formats described in RFC 1483.

Appendix IV. Ambiguity at the Edge [Informative]

IV.1 Ambiguous Encapsulation Information At The Egress MPC

In certain network configurations, traffic associated with two or more distinct flows of data can converge at a single node within the network and be expected to diverge again on leaving this node. In a typical store-and-forward device in a network, this would present no problems. The device would either make forwarding decisions based internetwork-layer information (e.g., in a router or layer-3 aware switch), or it would leave layer 2 headers intact (e.g., in a bridge).

Because data arriving on an MPOA shortcut VCC does not include an ISO layer 2 header, the impact of temporarily merging distinct data flows may result in a need for distinct cache impositions for each data flow. Merging of flows at the last-hop ATM/MPOA router (MPS) prior to the egress MPC, this router's next hop(s) or only at the egress MPC itself would result in the use of multiple distinct DLL headers for a given internetwork-layer destination. Because the egress MPC is required to prepend the correct DLL header to each data packet received on a shortcut VCC prior to forwarding it on an appropriate port, the egress MPC must be able to distinguish flows using more than the internetwork-layer destination.

IV.2 Resolving Egress Ambiguity

An egress MPC is able to detect when such an ambiguity occurs because it receives a new cache imposition (with a new cache ID) that has the same layer 3 destination and source ATM address as one of its existing cache entries, but a different next hop DLL header. At this point, an MPC implementation should assume that the ingress MPC (or NHC) will re-use a shortcut VCC associated with the existing cache entry. Therefore, the egress MPC must take some action to ensure that it will be able to distinguish packets arriving on such VCCs and make the correct association of cache and flow.

Assuming that the egress MPC is not an active router or layer 3 switch (i.e. it does not have co-resident MPS), the actions that it might take are:

- refuse the cache imposition (force the flow associated with the newer cache to continue to use default forwarding),
- return a distinct ATM address for the new cache imposition or
- assign a tag value in the cache acknowledgment.

The latter option may be used only if the cache imposition includes a TAG TLV indicating that the ingress is prepared to receive a tag in its response and use the tag for all frames transmitted on the shortcut. This is not the case, for example, if the "ingress" is a standard NHC.

IV.3 Ambiguity At The Ingress

The combination of ambiguous use of shortcut VCCs and data-plane purge results in the potential for ambiguous purge messages being received over a shortcut VCC. This will result either when an egress MPC is using tags to distinguish flows or when a single VCC is used to carry several flows and the VCC terminates at an ATM router (NHS or MPS) - which may or may not use tags to distinguish flows. In the worst case, the ingress MPC is forced to be conservative in purging cache entries and reissue MPOA Resolution Request(s) for the layer 3 destination address associated with the purge message received.

Appendix V MPOA-friendly NHRP implementations [Informative]

This appendix lists some optional extensions and procedures that an NHC/NHS may wish to implement for more efficient interoperation with MPOA devices. These features are not needed for interoperation between NHRP and MPOA devices, but in some cases efficiency may be improved.

1. support the use of the MPOA tag extension
2. support the use of the ATM Service Category extension
3. always include a non-zero value for MTU size in a Resolution Reply
4. support CPCS-SDU size negotiation during signalling

Note that the NHRP specification requires correct processing of a purge message received on any VCC; therefore, MPOA data plane purges will be handled correctly.

Appendix VI. MPOA Requirements for Co-Located LEC [Informative]

To support the MPOA device discovery mechanism described in Section 4.2, LECs must support the functionality described in this section that is not defined in [LANE].

VI.1 Support MPOA Device Type TLV Association

LANE allows the use of multiple ATM addresses by a LEC. The MPOA Device Type TLV is associated with an ATM address of the LEC and, therefore, all MAC addresses behind it. The LE_ASSOCIATE.request interface defined in LANE only supports association of TLVs with single LAN_Destination address. For an MPS in a router with a limited number of MAC addresses, it may be feasible to individually associate the MPOA device type TLV with each of these MAC addresses; however, for an MPC in a bridge that dynamically learns a large number of MAC addresses, it is not likely feasible to individually associate the MPOA device type TLV with each of these dynamically learned MAC addresses. To support the MPOA learning, a LEC should allow the MPOA component to associate the MPOA Device Type TLV with an ATM address and all of the LAN Destinations behind it.

To further facilitate learning, the MPOA Device Type TLV should be carried in all LE_ARP requests originating from a LEC that uses the ATM address associated with the MPOA Device Type TLV as the source ATM address in the LE_ARP.

VI.1 Support for LECs that do Source Learning

It is possible for LECs to learn MAC-to-ATM address mappings by observing data flowing over a LANE Data Direct VCC. When the LEC sees a new source MAC address, it can add the MAC-to-ATM address mapping to its LE_ARP cache. This type of source learning can defeat the MPOA device discovery process unless the MPOA Device Type TLV is associated with the learned MAC addresses properly.

To properly associate the Device Type TLV with learned MAC addresses, a LEC needs to abide by the following rules:

1. For each source ATM address from which a LEC is performing source learning, the LEC must issue at least one LE_ARP_REQUEST for a learned MAC address to determine the MPOA device type of the LEC with the ATM address.
2. The MPOA Device Type TLV received in an LE_ARP_RESPONSE must be associated with all MAC addresses subsequently learned to be associated with the ATM address in the LE_ARP_RESPONSE and not just the MAC address in the LE_ARP_RESPONSE.
3. An LEC that receives an indication of a change of MPOA device type for one MAC address must assume that this change effects all the MAC addresses learned from the same ATM address.

VI.1 Support for LLC Multiplexing

When LLC multiplexing is used, the MPOA device type is associated with the <ELAN-ID,LLC-MUXED-ADDRESS> pair, instead of just the ATM address as described above.