ATM Forum: Technical Committee

ATM Forum/ 95-0824r8

****

Source: Multiprotocol Sub-Working Group

Editor: Caralyn Brown
Bay Networks, Inc.
2 Federal Street
Billerica, MA 01821

Phone: (508) 436-3835
Fax: (508) 670-8760
Email: cbrown@baynetworks.com

****

Title: Baseline Text for MPOA

****

Date: June 21, 1996~~May 24, 1996~~

****

Distribution: Multiprotocol Sub-Working Group

****

Abstract: This contains the baseline text for the MPOA sub-working group as agreed to at the February technical committee meeting held in Beverly Hills, CA.

****

Notice: This contribution has been prepared to assist the ATM Forum. This document is offered to the Forum as a basis for discussion and is not a binding proposal on Bay Networks, Inc. The requirements are subject to change in form and value after more study.

****

0. Work Topic List **44**

1. Introduction **55**

1.1 State of this Document 55

2. Terms and Definitions. **55**

2.1 Definitions 55
2.2 Routing and Addressing Model Interactions 66
2.3 Addressing Models 66
2.4 Routing Models 77
2.5 Acronyms 77

3. MPOA Service Overview **77**

3.1 What is MPOA 77
3.2 Services Provided by MPOA 88
3.3 Services Required by MPOA 88

4. Elements of the MPOA Solution **98**

4.1 Overview 98
4.2 Logical Components of the MPOA Solution 109
4.3 Information Flows in the MPOA Solution 1110
4.3.1 Summary *1110*
4.3.2 Client to Server Flows *1211*
4.3.3 Server to Server Flows *1312*
4.3.4 Encapsulation *1312*
4.3.5 Control and Reliability *1312*
4.4 The EDFGs, however, which handle the internetwork data forwarding, are not intrinsically privy to changes in the routing topology. Thus, for some routing changes there will be a need to reliably notify EDFGs of the changes. RSFGs will provide these notifications. The exact circumstances this applies to, and the reliability mechanism to be used are still under study. Also under study are the time constraints required, particularly with regard to the prevention of forwarding loops. Cache aging ensures that even when the reliability mechanism fails, the system is self-correcting. It is currently felt that the time span for aging will be longer than is desirable for these situations. ATM Attached Host Models 1413

5. MPOA Service Specification **1413**

5.1 Summary 1413
5.2 Startup/Configuration Behavior 1514
5.2.1 Summary *1514*
5.2.2 RSFG/ RFFG Configuration *1514*
5.2.3 ICFG/DFFG Configuration *1615*
5.2.4 Edge Device Functional Group (EDFG) Configuration *1615*
5.2.5 ATM-Attached Host Functional Group (AHFG) Configuration *1716*
5.3 Registration 1716
5.3.1 Summary *1716*
5.3.2 ICFG Registration *1716*
5.3.3 AHFG Registration *1716*
5.4 MPOA Target Resolution 1817
5.4.1 Initiator Perspective *1817*
5.4.2 Authoritative Responder Perspective *1817*
5.4.3 Target EDFG Perspective *1817*
5.5 Data Transfer 1817
5.5.1 Default Unicast Data Transfer *1818*

5.5.2 Short Cut Unicast Data Transfer ..... 20~~18~~
5.5.3 Broadcast/Multicast Data Transfer ..... 20~~18~~
5.6 Spanning Tree Support ..... 20~~18~~
5.7 Replication Support ..... 20~~19~~
5.8 Multicast Support ..... 20~~19~~
5.8.1 Overview of ICFG.mars and mclient Interaction. ..... 21~~20~~
5.8.1.1 Message types for ICFG.mars interactions. ..... 22~~20~~
5.8.1.2 Registration. ..... 23~~21~~
5.8.1.3 Joining and Leaving Groups. ..... 23~~21~~
5.8.1.4 Providing Redundant ICFG.mars Entities. ..... 23~~22~~
5.8.2 Overview of the Data Path Management. ..... 24~~22~~
5.8.3 Overview of ICFG.mars and DFFG.mcs Interaction. ..... 24~~22~~
5.8.4 MPOA Multicast Related LLC/SNAP Code Points. ..... 25~~23~~
5.8.5 General Control Message Format. ..... 26~~24~~

6. Detailed Component Behaviors ..... **26~~24~~**

6.1 EDFG Behavior ..... 26~~24~~
6.1.1 EDFG Inbound Flow ..... 28~~26~~
6.1.2 EDFG Outbound Flow ..... 29~~27~~
6.1.3 Legacy to Legacy Flows ..... 29~~27~~
6.1.4 Cache Management ..... 30~~28~~
6.1.4.1 Ingress Cache Entry Management ..... 30~~28~~
6.1.4.2 Egress Cache Entry Management ..... 30~~28~~
6.2 ATM Host Behavior ..... 31~~29~~
6.2.1 AHFG Behavior ..... 31~~29~~
6.2.2 Dual Mode Host Behavior ..... 31~~29~~
6.3 ICFG/DFFG Behavior ..... 31~~29~~
6.3.1 ICFG Behavior ..... 31~~29~~
6.3.2 Intra-IASG Coordination ..... 32~~30~~
6.3.3 DFFG Behavior ..... 32~~30~~
6.4 RSFG/RFFG Behavior ..... 33~~31~~
6.5 Cache Management ..... 34~~32~~

7. Detailed Protocol ..... **34~~32~~**

8. Signaling Details ..... **40~~32~~**

9. References ..... **40~~32~~**

0. Work Topic List

1. Spanning tree - VLAN environments and boxes that bridge one protocol and route others - issues lurk.
2. Details of component behavior.
    - AHFG
    - general service description
    - ATM host behavior
    - RSFG/RFFG
        - RSFG/RFFG I/O relationship
    - ICFG/DFFG
    - cache management
    - EDFG (Mike See)
        - Wording of how to deal with EDFG looping back
        - abstraction for legacy ports
        - ATM to legacy forwarding behavior (elaboration needed)
        - Inclusion of TLV to accept tags.
    - Address motion 7 from 4/96. (Andre Fredette)
3. Protocol description (NHRP, MARS, LANE references)
    - enhancements for registration
    - enhancements for cache management
        - triggers
        - egress impositions
        - maintenance/purge
    - enhancements for piggy-backed MAC information
    - NHRP (Eric Gray)
        - Use of NHRP Fields
        - Vendor private extension TLVs (tag and encapsulation)
4. Day in the life of a packet (Norm Finn, Dan Townsend and others)
5. Configuration.
    - Failover - standby how do you find them? (James Watt)
    - Dynamics and resiliency of configuration.(James Watt)
    - what is to be configured
    - configuration validity and consistency checks.
    - definition of configuration database and management of same
6. Replication of ICFGs (includes ipmc)
    - load sharing (per protocol and per service area)
    - redundancy
    - fail-over
    - selection of ICFG (see failover and sharing etc.)
    - capabilities (load sharing, some can do things and some cannot)
    - coordination
7. Signaling. (Code points etc. UNI 3.x and 4.x considerations)
8. Multicast issues
9. QoS support/issues.
10. Filtering support/issues.
11. IASG labeling - Partitioning or what tells you how to remerge?
12. Discovery/registration of internal components.
13. Support of additional legacy ports - non-LAN ports. (interfaces at the edge of MPOA)
    - ~~Frame Relay~~
    - Serial lines
    - etc.

~~14. What is MPOA? (George Swallow)~~
14. What is the consequence of the RSctrl connection going down?
15. Create new drawing for figure 2 which addresses motions 10 and 11 from 4/96.
16. Handling of cache entries when the RSFG goes down. Including detection of this situation.

1. Introduction

1.1 State of this Document

Specific base choices were made to provide clear architecture. The detailed design choices in this document are open for modification.

Where applicable and equivalent we will re-use mechanisms from existing standards and not arbitrarily make alternate design choices. In particular, aspects of the LANE mechanisms, either current or prospective, may be applicable to the intra-IASG (Internetwork Address Sub-Group) protocols while aspects of the NHRP mechanisms may be applicable to the inter-IASG protocols.

It is acknowledged that this document may contain design choices that conflict with the foregoing statements and thus may need to be revised.

2. Terms and Definitions.

2.1 Definitions

Direct Set: A set of host interfaces which can establish direct layer two communications for unicast (not needed in MPOA).

Edge Device: A physical device which is capable of forwarding packets between legacy interfaces and ATM interfaces at any protocol layer, including, but not limited to, the internetwork layer It may or may not also participate in running one or more internetwork layer routing protocols.

Egress: A term used to describe the point where an outbound flow exits the MPOA system.

Emulated LAN: See ATM Forum LANE specification (ATM Forum document AF-0021).

Forwarding Description: The resolved mapping of an MPOA Target to a set of parameters used to set up an ATM connection on which to forward packets.

Host Apparent Address Sub Group:

A set of internetwork layer addresses which a host will directly resolve to lower layer addresses.

Inbound Flow: A term used to describe data flowing from the direction of a legacy port into the MPOA system.

Ingress: A term used to describe the point where an inbound flow enters the MPOA system.

Internetwork Address Sub-Group (IASG):

A range of internetwork layer addresses summarized in an internetwork layer routing protocol.

Internetwork Broadcast Sub-Group (IBSG):

A set of hosts which can receive a particular internetwork layer broadcast. Note that the scope of this is normally the same as IASG.

Internetwork Layer:	e.g. IP, IPX, DECnet routing, CLNP, AppleTalk DDP, IPv6, Vines
MPOA:	An effort taking place in the ATM Forum to standardize protocols for the purpose of running multiple network layer protocols over ATM.
MPOA Client:	A protocol entity which implements the client side of the MPOA protocol, , i.e. it exhibits either Edge Device Functional Group (EDFG) or ATM-attached Host Functional Group (AHFG) behavior.[1]
MPOA Server	A protocol entity which implements the server side of the MPOA protocol, i.e. it contains one or more of any of IASG Coordination Functional Groups (ICFGs), Route Server Functional Groups (RSFGs), Default Forwarder Functional Group (DFFG) or Remote Forwarder Functional Group (RFFG).
MPOA Service Area:	The collection of server functions and their clients . A collection of physical entities consisting of an MPOA server plus the set of clients served by that server.
MPOA Target:	A set of {protocol address, path attributes, e.g. internetwork layer QoS, other information derivable from received packet} describing the intended destination and its path attributes that MPOA devices may use as lookup keys.
One Hop Set:	A set of hosts which are one hop apart in terms of internetwork protocol TTLs (TTL=0 “on the wire”).
Outbound Flow:	A term used to describe data flowing from the direction of the MPOA system toward the legacy port.
Router:	A physical entity that is capable of forwarding packets based on internetwork-layer information, and which also participates in running one or more internetwork-layer routing protocols.
Routing Protocol:	A general term indicating a protocol run between routers and/or route servers in order to exchange information used to allow computation of routes. The result of the routing computation will be one or more forwarding descriptions.
Subnet:	The use of the term “subnet” to mean a LAN technology is a historical use and is not specific enough in the MPOA work. Please refer to Internetwork Address Sub-Group, Direct Set, SNDCF, Host Apparent Address Sub-Group and One Hop Set for more specific definitions.

2.2 Routing and Addressing Model Interactions

The following definitions of the addressing and routing models are included to assist in the understanding of various configurations. Any combination of routing and addressing models may be used in MPOA. Additionally, the selection of routing model is independent of the selection of addressing model.

2.3 Addressing Models

Peer Addressing Model: A model for mapping the addressing of an internetwork layer onto that used by the ATM fabric. It assumes that a sender can locally algorithmically translate the desired internetwork-layer destination address into the correct ATM address to use to place a call to that destination.

Separated Addressing Model: A model for mapping the addressing of an internetwork layer onto that used by the ATM fabric. It assumes that a sender must perform a dynamic lookup of the mapping from the desired internetwork-layer destination address to

[1] MPOA Clients and MPOA Servers may be co-located in a single physical entity.

the correct ATM address to use to place a call to the destination. This mapping information is provided by the destination or some proxy for it (e.g. network management or a network server) either a priori or else in response to a query. [2]

2.4 Routing Models

Integrated Routing Model: A model used to represent the internetwork-layer routing topology on top of an ATM topology. It integrates the internetwork-layer topology with that of the underlying ATM infrastructure into a single database of information.

Layered Routing Model: A model used to represent the internetwork-layer routing topology on top of an ATM topology. It separates information about the underlying ATM infrastructure from the layers above it, e.g. current utilization/congestion reports.[3]

2.5 Acronyms

AHFG	ATM-attached Host Functional Group
EDFG	Edge Device Functional Group
FG	Functional Group
DFFG	Default Forwarder Functional Group
IASG	Internetwork Address Sub-Group
IASGid	Internetwork Address Sub-Group Identifier
IBSG	Internetwork Broadcast Sub-Group
ICFG	IASG Coordination Function Group
MARS	Multicast Address Resolution Server
MCS	Multicast Server
MPOA	Multiprotocol Over ATM
NHRP	Next Hop Resolution Protocol
NHS	Next Hop Server
QoS	Quality of Service
RFFG	Remote Forwarder Functional Group
RSFG	Route Server Functional Group
SVC	Switched Virtual Channel

3. MPOA Service Overview

3.1 What is MPOA

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 layer 3 protocols on ATM. It provides direct connectivity between ATM attached devices (a.k.a. shortcuts) in order to reduce latency and the layer 3 processing load. MPOA provides a unified host behavior. MPOA is capable of using both routing and bridging information to locate the edge device closest to the addressed end station.

This framework is composed of a number of protocols being developed in both the ATM Forum and IETF. MPOA is adopting and integrating the NHRP work of the IETF and the LANE work of the ATM Forum. In some areas these protocols are being extend to better operate in the MPOA Framework. In some cases this work has occurred in the originating group, based on suggestions from MPOA. In other cases the MPOA group has made specific extensions to protocols.

[2] This is sometimes also known as “Overlay Addressing.”

[3] This is sometimes also known as the “Overlay Routing Model.”

Additionally the MPOA group is developing solutions to the areas of this framework not already addressed by other bodies. Most notable among this category is the separation of route calculation from layer 3 forwarding, a technique which has come to be know as a virtual router. This separation will provide three key benefits: to allow the integration of intelligent VLANs, to enable cost effective edge devices, and to provide a migration path for LANE devices.

~~**Editor's note: The following needs to be cleaned up in language and content using the bullet list included below.**~~

~~The Multi-Protocol Over ATM specification exists to solve two related problems.~~

~~The first is the need to integrate a number of current "solutions" for internetworking over ATM. The ATM Forum has developed the LANE specifications. The IETF is developing NHRP (Next Hop Resolution Protocol) and others. MPOA provides a framework in which these protocols may co-exist, and using which customers may migrate their solutions.~~

~~The second problem addressed by MPOA is the standardization of the protocols for supporting a virtual router over ATM. Specification is needed in this area to allow interoperability of components from different vendors.~~

- ~~Framework~~
- ~~Synthesis~~
  - ~~routing and bridging~~
  - ~~cut through~~
- ~~Virtual Router~~
  - ~~separation of routing protocol engine from forwarding engine~~
  - ~~interaction with virtualized (separated) VLANs~~

3.2 Services Provided by MPOA

The fundamental purpose of the MPOA Service is provide end-to-end internetworking layer connectivity across an ATM fabric, including the case where some internetworking layer hosts are attached directly to the ATM fabric, some are attached to legacy subnetwork technologies, and some are using ATM Forum LAN Emulation (LANE).

An internetwork can be sub-divided into one or more non-overlapping Internet Address Sub Groups (IASGs). An IASG is defined to be "a range of internetwork layer addresses summarized into internetwork layer routing." For the purposes of this document, the scope of an IASG is taken to be identical to the Internetwork Broadcast Sub-Group (IBSG) for each of the members of the IASG (this saves creating yet another term, or heaven forbid, yet another acronym). An IASG is internetwork-protocol specific, that is if a host operates two internetwork layer protocols, it is a member of, at least, two IASGs.

Within an MPOA System, an IASG is identified by an IASG Identifier (often shortened to IASGid). This identifier value is an opaque bit pattern of a length and format to be determined; the definition of values for and semantics of a small number of distinguished values, e.g. "black hole", "null", "foreign", is also a matter for further study.

3.3 Services Required by MPOA

MPOA makes use of ~~–~~ATM AAL5 for data transfer. It uses the capabilities of signaling as defined by the ATM Forum UNI 3.1, with options to make use of signaling 4.0 capabilities where they are available.

For use with internetworking protocols in deployments where routing is used, MPOA uses the routing capabilities of the individual underlying internetworking protocol.

For supporting distributed virtual LANs, MPOA makes use of the IEEE 802.1d protocol. It may be able to make use of 802.1q when that is fully specified.

MPOA Target resolution (section 5.4) depends on the existence of augmented NHS functionality at each MPOA service element.

[editor's note: This section is quite awkward and we need to make some wording changes.]

3.4 Future Considerations

3.4.1 Frame Relay

Support for RSFG/RFFG/DFFG use of frame relay is deferred to future phases of MPOA. There are several possible ways for an implementation to support frame relay in the interim.

Given frame relay SVC interworking with ATM one could extend the MPOA service transparently over the frame relay as well.

If one has frame relay SVCs and a service analogous to MPOA, an edge device might be serving an EDFG function simultaneously on ATM and frame relay.

If one is running bridging over frame relay, the edge device can participate in the frame relay bridging and from the point of view from MPOA this is indistinguishable from other bridging technologies.

4. Elements of the MPOA Solution

4.1 Overview

The MPOA solution presented in this document consists of a number of logical components and information flows between those components. The logical components are of two kinds: MPOA clients and MPOA servers – see section 4.2~~4.2~~ for details of the components and their categorization. The information flows between logical components, described in Section 4.3~~4.3~~ below, are carried over ATM SVCs using LLC/SNAP encapsulation.

MPOA servers exchange information in order to ensure that each server has appropriate, current information for responding to requests it services.

MPOA clients maintain local caches of information. These caches include mappings from MPOA target information to forwarding descriptions. There are several sets of such caches, used for different data flows. Information for the caches is generated by MPOA servers in response to requests or detected patterns in the default data flow. These cache entries are aged by the clients according to the lifetime specified by the MPOA server that provided it, i.e. ICFG (IASG Coordination Functional Group) or RSFG (Route Server Functional Group). The provider of a cache entry may "retract" it at any time, e.g. when routing changes or a station moves.

Replication of server components may be required for reasons of capacity and/or availability. The solution described herein explicitly admits and provides for the replication of server comments with one important caveat: clients are blissfully unaware of these mechanisms.

4.2 Logical Components of the MPOA Solution

For clarity, this document attempts to restrict itself to the discussion of functions performed by logical components and defers the composition of logical components into physical devices to those who develop products.

Functional Group (FG): A collection of "functions" related in such a way that they will be provided by a single logical component of MPOA. Examples include the Route Server Functional Group (RSFG) the IASG [Internetwork Address Sub-Group] Coordination Functional Group (ICFG), and the ATM-attached Host Functional Group (AHFG). One or more of any of these functional groups may co-reside in the same physical entity. MPOA allows arbitrary physical locations of these groups.

All of the Functional Groups present in an MPOA System are shown in Figure 11~~Figure 1~~ and described briefly below.

The *Edge Device Functional Group (EDFG)* is the group of those functions that are performed by a device which provides internetworking level connections between legacy internetworking subnetwork technologies and ATM. These functions are related to the forwarding of internetwork datagrams, and not to the operation of internetwork routing protocols.

The *ATM-attached Host Functional Group (AHFG)* –is the group of functions performed by an ATM-attached host that is participating in the MPOA service.

The *IASG Coordination Functional Group (ICFG)* is the group of functions performed to coordinate the physical distribution of a single IASG across multiple legacy LAN physical ports and/or ATM stations, e.g. permitting the distribution of a single IPv4 subnet across multiple Ethernet segments whilst maintaining correct operation. It is concerned with managing the information about members of the IASG, including ATM address, and for AHFGs, the assigned MAC address. This registration information includes participation in internetwork level multicasting, based on the Multicast Address Resolution Server (MARS) specification from the IETF.
[editor's note: this seems awkward...we should work on this]

The *Default Forwarder Function Group (DFFG)* is the group of functions performed in association with forwarding traffic to/from AHFGs within an IASG in the absence of direct client to client connectivity. The DFFG may also forward traffic to other forwarders when the traffic is destined outside the IASG in the absence of direct client to forwarding server connectivity. The DFFG also acts as the Multicast Server (MCS) in an MPOA based MARS implementation.

The *Route Server –Functional Group (RSFG)* is the group of functions performed to provide internetworking level functions in an MPOA System. This includes running conventional internetworking Routing Protocols and providing inter-IASG destination resolution.

The *Remote Forwarder Functional Group (RFFG)* is the group of functions performed in association with forwarding traffic from one IASG to another, from one MPOA client in one IASG to another IASG, from one IASG to the destined MPOA client in another IASG or from an MPOA client in one IASG to an MPOA client in another IASG.

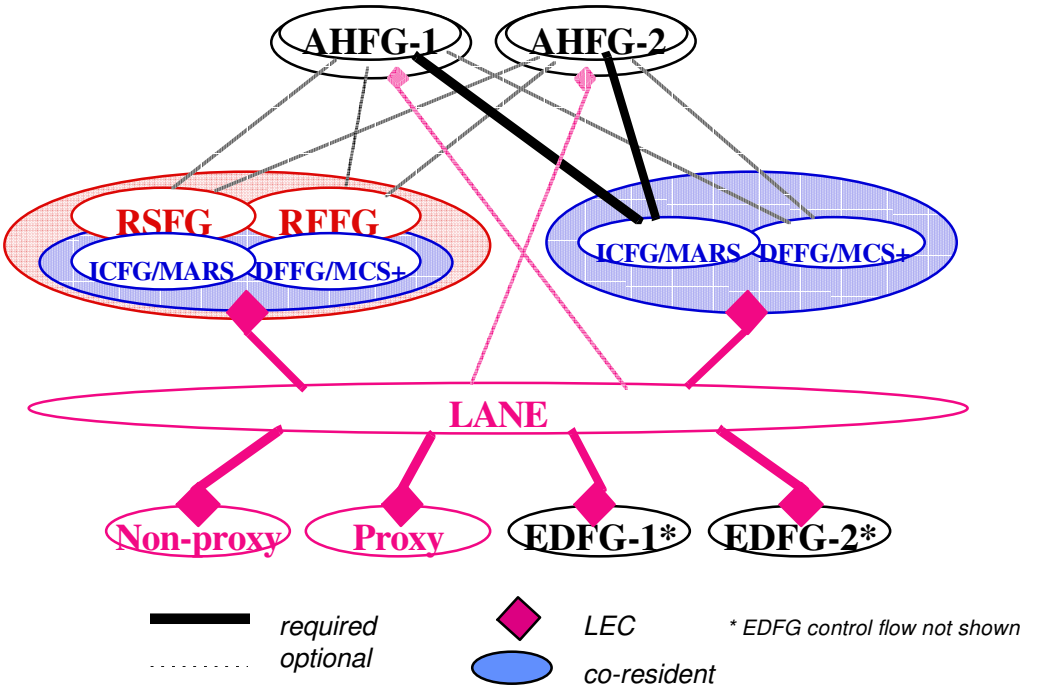


Figure 1~~1~~ - The Function Groups in an MPOA System

4.3 Information Flows in the MPOA Solution

4.3.1 Summary

The MPOA solution involves many information flows which can be categorized as follows:

a) configuration flows – all functional groups use such a flow to retrieve configuration information,

b) data transfer flows – the end goal of the system (note that communications between an RSFG and a router appear as data transfer to the rest of the system)

c) client-server control flows – used by clients to inform and query the MPOA servers and

d) server to server flows – used by servers to provide the illusion of a single service whilst distributing the service across multiple devices for reasons of capacity and/or availability. Figure 22~~Figure 2~~ shows the client-server and server-server information flows. The data transfer and configuration flows are suppressed to keep the diagram readable.

Section 4.3.4~~4.3.4~~ discusses the encapsulation of data within a flow and Section 4.3.5~~4.3.5~~ discusses issues with the reliable transmission of control and data frames. More detailed discussion of the information flows is contained in the area descriptions in Section 5.2~~5.2~~ through 5.6~~5.6~~.

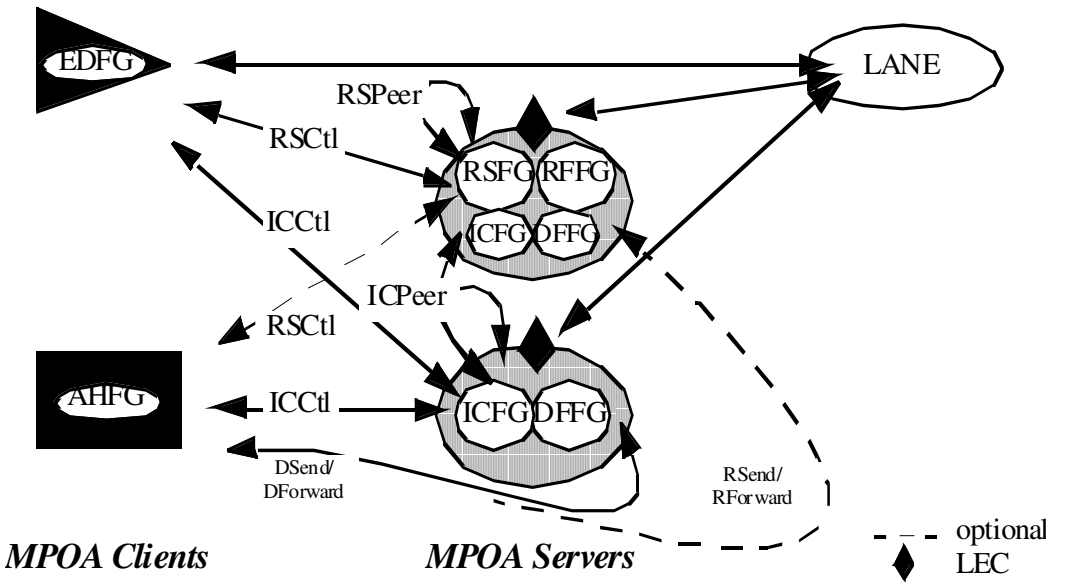


Figure 22 - Information Flows in an MPOA System

editor's note: motions 10 and 11 from 4/96 indicated that we needed to put additional flows into this diagram. Because of the chaotic state of this diagram, it might be better to create new drawings. An item has been added to the to do list of section 0.

Details of SVC establishment and detailed header and control packet formats are both matters requiring further study. The detailed data formats of user data are specified as part of the internetworking protocol used and thus need not be specified as part of the MPOA solution. Issues associated with SVC setup include, but are not limited to: aging of VCCs, provision and use of QoS information, selection of signaling Information Elements (IEs), and response to disconnects.

The *Configuration* information flows are not shown in Figure 22. Each functional group of the MPOA solution, client and server alike, creates such a virtual circuit at startup and uses it to retrieve the necessary configuration information. Configuration is discussed in Section 5.2.

Although the *Data Transfer* information flows (DSend/DForward and RSend/RForward) are shown in Figure 22 above their description is deferred until section 5.5.

4.3.2 Client to Server Flows

There are two kinds of client-server flows in the MPOA solution: RSFG Control (RSCtl), and ICFG Control (ICCtl).

RSCtl information flow is used by MPOA clients and ICFGs to obtain information from the RSFG when trying to resolve the destination internetwork address to an ATM address for inter-IASG data transfer.

editor's note: incorporate the additional arrows from contributions 507 and 506. Perhaps we should just use the drawings from these contributions.

ICCtl information flow is used by MPOA clients to obtain information from the ICFG in support of destination resolution in the intra-IASG case.

[Editor's note: the following sections were removed as a result of contribution 96-0559. It seems that they should be preserved somewhere. How about section 5.5 as suggested at the bottom of section 4.3.1? For now I'm leaving them here]

DSend information flow is used by MPOA clients and servers to transmit data frames in the absence of a direct, client-to-client, short-cut connection and for broadcast and multicast frames. Broadcast and multicast data transfer is described in Section 5.5.2 and unicast data transfer is described in Section 5.5.1.

DForward information flow is used by the DFFG to forward frames to at least those MPOA clients and servers, within the IASG, addressed by the frame.

RSend information flow may be used by MPOA clients (and potentially the DFFG) to forward frames out of an IASG while RForward information flow is used to forward frames into an IASG in the absence of direct IASG to IASG connectivity.

4.3.3 Server to Server Flows

There are two distinct kinds of server to server information flows in the MPOA solution: RSFG to RSFG (*RSPeer*) flows, ICFG to ICFG (*ICPeer*) flows. These flows are labeled in Figure 22~~Figure 2~~.

The *RSPeer* information flow is used by an RSFG to forward Destination Resolution queries (and to receive the response) for destinations in an IASG that is not served by the original RSFG. This flow may also be used for the change of conventional routing information.

The *ICPeer* information flow is used by an ICFG to distribute topology information to all of the ICFGs that serve a given IASG. The types of connections (point-to-multipoint or point-to-point) and the protocol to be used is still under discussion. It is expected that this protocol will be the same as that used by other services for this same coordination problem.

4.3.4 Encapsulation

By default, MPOA uses LLC~~/SNAP~~ encapsulation for all information flows. ~~–~~This is used in accordance with the rules defined in RFC 1483. The call setup mechanism provides for the negotiation of traffic encapsulations. This mechanism is used by MPOA (the exact details still require specification) and, therefore, allows functional groups to bi-laterally agree on alternative encapsulations. This negotiation may result in no explicit header being used over VCs, with the header being implicit in the usage of the VC.

To ensure interoperability, all MPOA devices are required to be able to use LLC~~/SNAP~~ encapsulation for all information flows on all VCs. This includes VCs for which the functional group "knows" that no actual sharing will occur. Negotiation of other encapsulations for MPOA flows is always optional.

The LLC/SNAP code points used by the MPOA solution require specification but will include any existing code points that can be re-used. New code points, when needed, will be defined using the ATM Forum's OUI. [It is currently under study whether the query-response protocol between an RSFG and its locally served EDFGs will use the NHRP code points, or will need a new LLC/SNAP code point.]

4.3.5 Control and Reliability

The data transfer methodology of the MPOA system is intended to support unreliable internetworking protocols. Therefore, data transfer need not be augmented with any reliability protocol for operation over ATM.

However, forwarding information is being cached in the EDFGs, while it is maintained cooperatively by the ICFGs, RSFGs and, for legacy stations, by the EDFGs themselves. When this information is incorrect, traffic may be sent to the wrong destinations, possibly resulting in routing loops or excessive data loss.

EDFGs may locally detect some of these errors, such as the motion of a legacy station, within appropriate times. This may require notification of certain errors to the ICFG, or to stations using cut-through VCs terminating on an EDFG. No special reliability is required for this, since a persistent problem will result in persistent notifications. For LANE forwarding, the LANE mechanisms will be used. [**editor's note: this paragraph is very awkward and should be modified at the editing session]**

The EDFGs, however, which handle the internetwork data forwarding, are not intrinsically privy to changes in the routing topology. Thus, for some routing changes there will be a need to reliably notify EDFGs of the changes. RSFGs will provide these notifications. The exact circumstances this applies to, and the reliability mechanism to be used are still under study. Also under study are the time constraints required, particularly with regard to the prevention of forwarding loops. Cache aging ensures that even when the reliability mechanism fails, the system is self-correcting. It is currently felt that the time span for aging will be longer than is desirable for these situations.

4.4 ATM Attached Host Models

editor's note: further contributions are required to fill in the text of this section. Please align contributions with the motions that were passed during the December meeting.

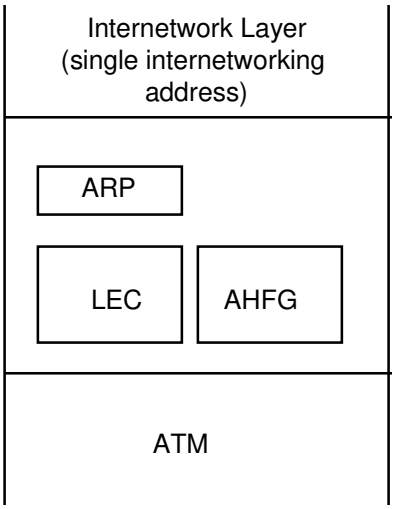


Figure 33 - Dual Mode Host

5. MPOA Service Specification

5.1 Summary

An MPOA System, which is comprised of the functional groups described in Section 4.24.2, covers the following areas:

a) *Configuration*: this ensures all functional groups have the appropriate set of administrative information;

b) *Registration and Discovery* : this includes the functional groups informing each other of their existence and of the identities of attached devices;

c) *Destination Resolution*: this is the action of determining the route description given a destination Internetwork Layer address (and possibly other information, e.g. QoS). This is the part of the MPOA System that allows it to perform "cut through" (with respect to IASG boundaries);

d) *Data Transfer:* is getting internetworking layer data from one MPOA client to another;

e) *Intra-IASG Coordination*: is the function that allows IASGs to be spread across multiple physical interfaces;

f) *Routing Protocol Support*: this allows and MPOA System to interact with conventional internetworks;

g) *Spanning Tree Support*: this allows the MPOA System to interact correctly with existing extended LANs;

h) *Replication Support:* this provides for replication of key components for reasons of capacity or availability.

Sections 5.25.2 through 5.75.7 provide more detailed descriptions of each of these areas.

As discussed in Section 4.14.1, the system functions as a set of functional groups that exchange information in order to exhibit the desired behavior(s). To provide an overview of the system's operation, the behavior of the components is described in a sequence order by "significant events."

In the descriptions which follow, all externally visible information flows in the MPOA System are assumed to be carried in VCCs over the ATM fabric. It may be the case, that for some compositions of the functional groups into physical devices, ATM-level communication between certain functional groups is unnecessary.

5.2 Startup/Configuration Behavior

5.2.1 Summary

All functional groups will have/know/be configured with a way to contact an appropriate configuration server. The number of servers, synchronization of server databases, and related tasks need to be specified, however it is expected the approach will closely follow, if not simply re-use, the mechanisms specified for the LAN Emulation Configuration Service (LECS). Note that the rest of this document assumes that all of the configurations delivered to the various components are consistent across the system.

In the following, the assertion that the configuration server returns a specific RSFG, or ICFG address does not mean there is only one. It merely means that a given station only needs to talk to one of each kind of server for a given IASG. An alternate design would return the full list of alternate servers to the client, and let the client chose.

The configuration information specifies that various functional groups are handed full lists of the members of other functional groups. This is to ensure that appropriate queries can be generated, and/or that appropriate responses are generated to such queries (for example, an RSFG gives supplemental information to an EDFG that it does not provide to an AHFG). Whether this information is actually provided by configuration, or is learned through some other mechanism is still under study.

5.2.2 RSFG/ RFFG Configuration

When it starts up, an RSFG/RFFG will get the following configuration information:

a) A list of IASGs it is providing service for, and for each IASG:

i) the identifier assigned to the IASG and

ii) the internetwork protocol for the IASG,

iii) the RSFG's internetwork address on the IASG,

iv) the information needed to join an ELAN, if any, for this IASG;

v) the ATM addresses of the EDFGs if required (see section TBD.)[4]

b) A list of routing protocols to use, and the normal configuration information required for each, including interfaces and peers, as defined by the specification of the routing protocol.

The RSFG/RFFG will establish communication with the ELAN, if any, associated with the IASG.

It will then begin operating the routing protocols according to the configuration information. Actual transmission of routing information will be over the lower layer protocol according to the rules for the protocol family.

5.2.3 ICFG/DFFG Configuration

When it starts up, an ICFG/DFFG will get the following configuration information:

a) A list of IASGs it is providing coordination for, and for each IASG:

i) the identifier assigned to the IASG and

ii) the internetwork protocol for the IASG,

iii) Other ICFGs to perform coordination with in support of this IASG.

iv) the information needed to join an ELAN, if any, for this IASG,

v) the address(es) for the RSFG/RFFG pair to use for internetwork forwarding,

vi) the information needed to generate MAC addresses on behalf of AHFGs when necessary;

vii) the ATM addresses of the EDFGs if required (see section TBD.)[5]

All ICFGs serving an IASG know about all others as this avoids single-points of failure or distinguished behaviors.

The ICFG will establish VCs to each distinct member of the list of peer ICFGs. Note that if the same ATM Address appears in several lists of peer ICFGs, only one VC is needed between the two servers. The ICFG/DFFG will establish communication with the ELAN, if any, associated with the IASG.

5.2.4 Edge Device Functional Group (EDFG) Configuration

When it starts up, each EDFG will get the following configuration information:

a) A list of IASGs it is to be a member of, and for each IASG:

i) The IASG identifier,

ii) The internetwork protocol supported,

iii) the ATM addresses of the RSFGs, RFFGs, ICFGs and DFFGs if required (see section TBD.)[6]

iv) the MTU for the IASG.[7]

v) the information needed to join an ELAN

b) For each legacy port, a list of IASGs supported on the port;

---

[4] Editor's note: Need contribution for different cache creation initiation methods, and their supporting information requirements.

[5] Editor's note: Need contribution for different cache creation initiation methods, and their supporting information requirements.

[6] Editor's note: Need contribution for different cache creation initiation methods, and their supporting information requirements.

[7] This is for further study. Contributions are requested.

c) The ELAN(s) to talk to for spanning tree propagation.

The EDFG will establish RSCtl and ICCtl information flows, if appropriate and will establish communication with the ELAN associated with each IASG.

The spanning tree ELAN(s)[8] is used to ensure that an entire MPOA family is participating in a single set of Spanning Tree exchanges.

5.2.5 ATM-Attached Host Functional Group (AHFG) Configuration

When it starts up, each host will get the following configuration information:

a) A list of IASGs it is to be a member of, and for each IASG:

i) The IASG identifier,

ii) The internetwork protocol supported,

iii) an ICFG to use for the IASG;

iv) the MTU for the IASG.

The host will establish VCs to each distinct ICFG. After the VCCs are setup, the host will proceed with registration as described in Section 5.3~~5.3~~ below. Note that use of a protocol-specific auto-configuration capability may require configuration information to be provided to the AHFG beyond that listed above. **Note: whether information for coordinating AHFG and LANE requires further thought. Contributions are requested.**

5.3 Registration

5.3.1 Summary

Registration in the MPOA solution is the set of exchanges that are used by the functional groups to inform each other of their existence, capabilities and interests. This mechanism is used by the MPOA solution, along with routing protocol support to ensure that the server components have accurate knowledge of the topology of the directly-attached network both at the MAC and internetwork levels.

5.3.2 ICFG Registration

The ICPeer flow is used for the coordination of information among ICFGs serving an IASG. The initial process of detecting a neighbor, establishing an adjacency with him, and determining what information is to be exchanged, can be thought of as a registration process. However, it is described as part of the synchronization protocol.

5.3.3 AHFG Registration

As the appropriate VCCs become available, each AHFG will register itself with each distinct ICFG listed in its configuration. The AHFG will provide a list of (IASGid, internetwork address, and optionally MAC address) tuples to the ICFG. The AHFG must also inform the ICFG whether or not to serve as the AHFG's proxy on an ELAN. If the AHFG does not provide a MAC address and the ICFG has been requested to proxy for the AHFG on an ELAN, the ICFG will provide a MAC address for the AHFG.

When an ICFG receives an AHFG registration, it propagates the information to ICFGs using the ICPeer flow. The ICFG will also return to the AHFG the ATM address of the DFFG to use. That DFFG will be performing the proxy for the AHFG on the ELAN.

[8] Contributions are requested to clarify the spanning tree propagation.

5.4 MPOA Target Resolution

MPOA target resolution uses an extended NHRP query response protocol to allow clients to determine the ATM address for end points of short cut VCs. In the following we describe the protocol from the perspective of the initiator of a query, the last RSFG to handle the query and the actual target.

5.4.1 Initiator Perspective

When an AHFG wishes to send traffic to an internetwork destination, it may send the traffic to the DFFG for forwarding. However, if a significant amount of information is to be sent, whether intra-IASG or inter-IASG, the AHFG needs to determine the ATM Address to send data to. This is done by sending a query to the ICFG using the ICCtl flow. Upon resolving the query, the ICFG responds to the AHFG with the correct ATM address. ~~–~~If the AHFG uses mask and match behavior, it may send appropriate queries (e.g. those for destinations outside the IASG) directly to an RSFG using the RSCtl flow.

Editor's note: EDFG perspective needed.

In the event that more than one flow exists on which an NHRP reply may be returned, an MPOA Client (EDFG or AHFG) must be prepared to receive the reply on any of these flows.

5.4.2 Authoritative Responder Perspective

When an NHRP request targeted for an AHFG arrives at the authoritative RSFG/ICFG, the RSFG/ICFG generates a response.

A message, defined in section (**need reference to format section here**) below, may be generated by the MPOA service component providing the authoritative reply to the NHRP request and forwarded to the MPOA target client. This message is part of a request/response (assert/acknowledge) protocol that serves a dual purpose - it provides required encapsulation information to an EDFG such that the EDFG may prefix legacy bound ATM traffic and it verifies that the MPOA Client can accept an incoming flow. The MPOA service component originating this message may wait for a reply before returning the corresponding NHRP reply message to the NHRP request originator. Because encapsulation information is required before shortcut traffic may be forwarded to a legacy interface, this message MUST be provided if the MPOA target client is an EDFG.
Editor's note: need clarity on beginning of paragraph and tag propagation.
Editor's note: Add text for NHRP triggers (some where but probably not in this section)

5.4.3 Target EDFG Perspective

Editor's note: This should describe reception and response to egress cache imposition, including tag generation.

5.5 Data Transfer

(editor's note: This section still needs some work.)
Unicast data flow through the MPOA system has two primary modes of operation: the default flow and the shortcut flow. Shortcut flows are established by the cache management mechanisms. The default flow mechanism provides for data forwarding when shortcuts do not exist.

5.5.1 Default Unicast Data Transfer

In the default situation the data frame is sent by an MPOA client to an RFFG or DFFG and then forwarded toward the final destination.

When an AHFG originates the packet, it may detect and choose, for internetworking destinations outside the IASG, to send the packet to the RFFG. Otherwise the AHFG sends the packet to the DFFG.

When an EDFG sends the packet in the default situation, it uses LANE for delivery based on the MAC address in the packet.

For a dual mode host originating data, it will use LANE for transmission of intra-IASG traffic and use the AHFG for inter-IASG traffic.

When a packet arrives at an RFFG for an inter-IASG destination it is forwarded by the routing system. When a packet arrives at a DFFG for a registered AHFG, the internetwork protocol packet is forwarded directly to the AHFG.

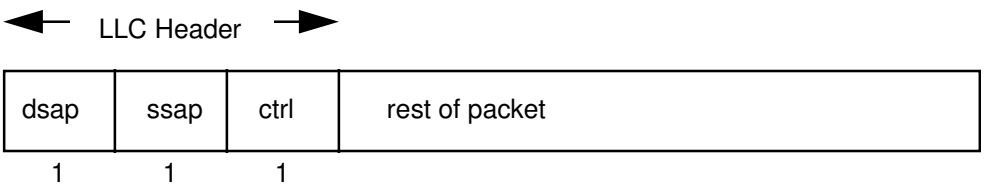
When a packet arrives at a DFFG from an AHFG for an intra-IASG destination that is not an AHFG, the DFFG will add the appropriate MAC header and use LANE for further forwarding. If the DFFG does not have enough information to build the MAC header, appropriate address resolution is used to get the MAC information.

When a packet arrives at a DFFG for a destination which is not within the IASG, the packet is forwarded to an RFFG.

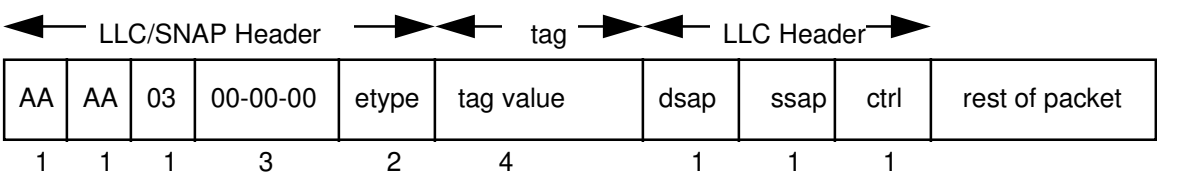
5.5.2 Data Transfer with Tags

All packets sent on a short-cut VCC will be LLC encapsulated. There may also be a mix of tagged and non-tagged packets on a VCC. The proposed encapsulation is for tagged packets to have an extra 12 bytes prepended to the non-tagged version of the same packet. This consists of an 8-byte LLC/SNAP header and the 4-byte tag field itself. After this comes the packet as it would have been sent if tags were not used. There will thus be a second LLC field in the packet.

Non-tagged packet:



Tagged packet:



The Ethertype value is TBD.

5.5.3 Default Server-Server Flow

In order to support the timely default forwarding of inter-IASG traffic, there must exist a set of flows among the DFFG(s) and RFFG(s) serving a given IASG. These flows are used by the DFFG(s) to forward non-

local data to the RFFG(s) for hop-by-hop forwarding, and are used by the RFFG(s) to deliver data to the correct DFFG for forwarding to destination AHFG(s) within an IASG.

In order to maximize technology re-use, and to ensure the reliable existence of these flows, these flows are actually transmitted via LANE, not via MPOA internetworking VCs.

5.5.4 Shortcut Unicast Data Transfer

When an MPOA client has an internetwork protocol packet to send for which it has an internetworking shortcut, the packet is sent over the shortcut VC with the appropriate internetworking encapsulation.

5.5.5 Broadcast/Multicast Data Transfer

When an AHFG wishes to send a broadcast or multicast frame, it uses the ICFG in a manner similar to the above. Specifically, it sends a query to the ICFG and gets an ATM address.

Inter-IASG broadcast/multicast transfer will presumably be done either by selective splicing onto the point-to-multipoint connection that will be used by the above, or by having another kind of server within the IASG accept the information and forward it to a different point-to-multipoint circuit for inter-IASG transfer.

For EDFGs forwarding broadcast/multicast, the same set of transmissions is required.

5.6 Spanning Tree Support

Just to re-iterate from a previous contribution, each EDFG must be running IEEE 802.1(d) spanning tree on each port. There is a specific IASG just for propagating spanning tree BPDUs among the EDFGs. Each EDFG is responsible for the generation of BPDUs into that IASG, and onto its attached legacy media.

Each EDFG will also notify the ICFG when any of its ports enter the Blocking or Forwarding state. (Note that the "ports" for this purpose include the ATM port, which can become blocked.) The blocking notification is critical since that results in invalidating routing/location information. The ATM Port blocking also interacts with "knowing" where things are, since other EDFGs are not directed to send intra-IASG traffic to that EDFG. The forwarding state notification serves to counterbalance the blocking notification (thus avoiding "assumptions"). –In addition, it permits the ICFG to determine if an entire IASG has gone blocked for access from ATM to Legacy. It is for further study whether the ICFG should then re-enable one port for routing, or whether sufficient information should be passed to the RSFGs to permit them to do so. (In the ICFG case, although there are multiple ICFGs, as long as they all come to the same conclusion about which port of which EDFG should be partially re-enabled, only one ICFG will actually do anything.)

The question whether distinct spanning trees are required to support both IEEE 802.1(d) Spanning Tree and Token Ring Spanning Tree simultaneously or whether a single spanning tree will suffice requires further study.

5.7 Replication Support

[editor's note: This was specified in the summary section, but no text has been suggested]

5.8 Multicast Support

The first version of the MPOA multicast service will focus on providing specific support for multicasting among MPOA clients that are directly attached to an ATM network. Furthermore, the first version will

specifically support only intra-IASG multicasting. Inter-IASG multicasting will initially utilize conventional layer 3 inter domain multicasting mechanisms.

An alternative expression of this focus is that it covers the propagation of layer 3 multicast traffic among EDFGs and AHFGs belonging to the same IASG. The MPOA multicast service model assumes that within EDFGs and AHFGs there will be components specifically designed to support the multicasting needs of the functional group as a whole. Further, the MPOA multicast service model assumes that ICFGs will contain components specifically designed to support the multicasting needs of the IASG they serve.

The following derived terms are introduced to identify these components of existing functional groups:

ICFG.mars
Those elements of the ICFG that support the address mapping and group membership traffic requirements of the MPOA multicasting service. Based on the IETF's Multicast Address Resolution Server (MARS) [1].

AHFG.mclient
Those elements of the AHFG that utilize the ICFG.mars to establish and manage ATM resources for intra-IASG multicasting.

EDFG.mclient
Those elements of the EDFG that utilize the ICFG.mars to establish and manage ATM resources for intra-IASG multicasting.

DFFG.mcs
An entity within the MPOA multicast service that behaves as a Multicast Server. It has no functions relating to unicast traffic.

The term 'cluster' will be used to indicate the subset of MPOA clients within an IASG that are directly ATM attached. A cluster may contain any combination of AHFGs and EDFGs.[9]

The short-hand term 'mclients' will be used to generically refer to AHFG.mclients and/or EDFG.mclients in the following sections.

Finally, in this first version of the MPOA multicast service it will be assumed that only one ICFG.mars is actively supporting a given IASG at any time (although multiple ICFG.mars' may be monitoring the IASG activity and be able to take over as backup if the active one fails).

5.8.1 Overview of ICFG.mars and mclient Interaction.

An ICFG.mars serves one or more mclients. Within a given IASG the relationship between the ICFG.mars and its mclients is shown in Figure 44~~Figure 4~~. Each mclient establishes its own bi-directional point-to-point VC to the ICFG when it needs to query the ICFG.mars or register/deregister a group membership for itself. This VC may be transient, and represents an instance of the ICFG Control (ICCtl) VC identified in section 4.3.1~~4.3.1~~.

To support the requirement that mclients are updated quickly with group membership changes that occur in other mclients within the cluster, the ICFG.mars manages a permanent, unidirectional point-to-multipoint VC, ClusterControlVC [*for want of a better name*], out to all cluster members.

[9] Editor's note: The term cluster is discussed in the Multicast appendix which follows this section. We need to work on the order of the document and/or define cluster elsewhere.

ICFG.mars messages on ICCtl and ClusterControlVC are LLC/SNAP encapsulated. They use a unique code-point allocated by the IETF to the MARS protocol. Other protocols may share these control paths between the ICFG and EDFG/AHFGs (e.g. unicast MPOA Target resolution based on NHRP) by using different LLC/SNAP code points.

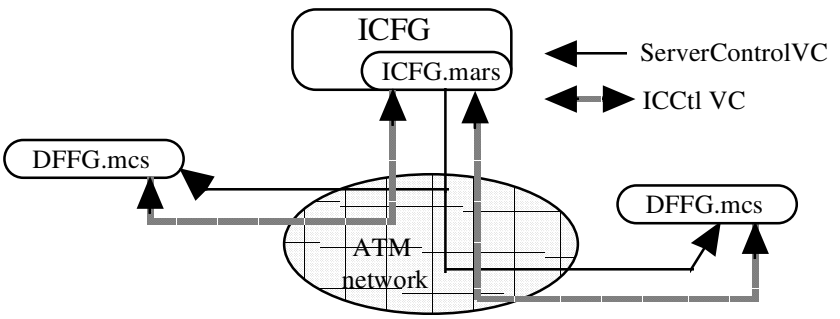


Figure 44 - Control paths between ICFG.mars and mclients.

All ICFG.mars messages transmitted on ClusterControlVC carry a 32 bit sequence number (Cluster Sequence Number, or CSN). This is incremented for every transmission on ClusterControlVC, and is tracked by mclients to allow them to detect if they've missed a previous general transmission from the ICFG.mars.

5.8.1.1 Message types for ICFG.mars interactions.

The following message types are taken from the MARS protocol. For convenience the IETF mnemonics will be adopted.

MARS_JOIN,
MARS_LEAVE.
Allow endpoints to join and leave specific layer 3 multicast groups. Also used to pass on to the entire cluster asynchronous updates of a group membership change.

MARS_REQUEST.
Allow endpoints to request the current membership list of a layer 3 multicast groups.

MARS_MULTI.
Allows multiple ATM addresses to be returned by the ICFG.mars in response to a single MARS_REQUEST.

MARS_NAK.
Explicit negative response returned by the ICFG.mars if no information is available to satisfy a previous MARS_REQUEST.

MARS_REDIRECT_MAP.
Allow ICFG.mars to specify a set of ATM addresses of backup ICFG.mars entities for its mclients.

MARS_GROUPLIST_REQUEST,
MARS_GROUPLIST_REPLY.
Allow ICFG.mars to indicate which groups have actual layer 3 members. (May be used to support higher layer functions such as IGMP in IPv4 multicast routing.)

Each message carries a 56 bit protocol identifier and an arbitrary byte string (up to 255 bytes long) representing layer 3 protocol multicast group address(es). The first 16 bits of the protocol identifier is an unsigned integer value taken from 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.

If the NLPID value of 0x80 is specified, the actual protocol being carried is identified by a SNAP value encoded in the remaining 40 bits of the protocol identifier field. Otherwise, the protocol is entirely specified by the initial 16 bit value.(As examples, a protocol identifier of 0x800 would indicate IPv4, and its multicast addresses would be 4 octets long. A protocol identifier of 0x86DD would indicate IPv6, and its multicast addresses would be 16 octets long.)

The message format is described in more detail in section 5.8.5~~5.8.5~~.

5.8.1.2 Registration.

In order to be added to ClusterControlVC, and so utilize an ICFG.mars, mclients must register with the ICFG.mars for their IASG. This is achieved by exchanging with the ICFG.mars a special form of the MARS_JOIN message on ICCtl VC. Registration also results in the ICFG.mars allocating each mclient a 16 bit ClusterMemberID (CMI, valid only for the time that the mclient remains an active leaf on ClusterControlVC).

A special form of the MARS_LEAVE message is used to deregister a cluster member.

5.8.1.3 Joining and Leaving Groups.

An mclient is a 'group member' (in the sense that it receives packets directed at a given layer 3 multicast group) when its ATM address appears in the ICFG.mars' mapping table entry for a given group's multicast address. It adds its ATM address to the membership map of a given group by sending the MARS_JOIN message on ICCtl VC.

Conversely, to leave a group an mclient needs to remove its ATM address from the ICFG.mars' mapping table for the given group. This is achieved by sending the MARS_LEAVE message on ICCtl VC.

When the ICFG.mars determines that all cluster members need to be informed of a change in group membership it propagates an appropriate MARS_JOIN or MARS_LEAVE on ClusterControlVC.

5.8.1.4 Providing Redundant ICFG.mars Entities.

Every mclient keeps a table of one or more ATM addresses representing the ICFG.mars' it should attempt to register with. This table is sorted in order of preference, and is stepped through sequentially during registration until registration succeeds. This table is also stepped through sequentially if re-registration occurs due to failure of the current ICFG.mars. As a minimum, each mclient must have at least one ATM address in this table.

The table is dynamically updated by the contents of MARS_REDIRECT_MAP messages that are regularly transmitted on ClusterControlVC. These messages contain a list of ATM addresses representing the set of ICFG.mars' that the current ICFG.mars considers to be its backups. When a client receives a MARS_REDIRECT_MAP it copies its contents into its local ICFG.mars table.

Hot swapping of mclients from one ICFG.mars to another is also triggered by MARS_REDIRECT_MAP. These procedures are discussed further in later sections.

5.8.2 Overview of the Data Path Management.

When an AHFG or EDFG requires an outbound path to the members of a layer 3 multicast group, and no such path pre-exists, its mclient constructs a MARS_REQUEST (representing the multicast MPOA Target) and sends it on ICCtl to the ICFG.mars (re-establishing ICCtl too, if necessary).

This query elicits a multicast Forwarding Description from the ICFG.mars in the form of one or more MARS_MULTI messages. The multicast Forwarding Description contains a set of one or more ATM addresses, which the mclient uses to specify the leaf nodes of a point-to-multipoint VC. The mclient then considers this outgoing VC to be the appropriate forward path for multicast traffic to the layer 3 group identified in the original MARS_REQUEST.

If no members exist for the group identified in the MARS_REQUEST, the ICFG.mars returns a MARS_NAK instead.

Once a forward path has been established, the mclient monitors incoming MARS_JOINs and MARS_LEAVEs. These will arrive on ClusterControlVC when the ICFG.mars is informing the entire cluster of a new member's arrival or an old member's demise. If a MARS_JOIN arrives indicating a new member on a group for which the mclient has an active forward path too, a new leaf node is immediately added to the forward path. If a MARS_LEAVE arrives indicating the loss of a member on a group for which the mclient has an active forward path too, the associated leaf node is dropped.

The mclient makes no assumption that the ATM addresses provided by the ICFG.mars in MARS_MULTI or MARS_JOIN messages actually represent the group members. The only assumption is that they represent the forward path determined by the ICFG.mars for that group. This allows the ICFG.mars to introduce multicast servers (DFFG.mcses) into the forward path in a manner transparent to the mclients. If a layer 3 group is being served by an DFFG.mcs then the ICFG.mars returns the ATM address of the DFFG.mcs in the MARS_MULTI. When a new group member joins or leaves an DFFG.mcs supported group, the ICFG.mars informs the DFFG.mcs rather than the mclients.

Data packets transmitted on the forward path are LLC/SNAP encapsulated. *For the purpose of this first iteration* they shall use the IETF's extended encapsulation for multicast packets, which carries within it the CMI allocated during mclient registration. This allows each mclient to detect reflected AAL_SDUs in a layer 3 protocol independent manner.

5.8.3 Overview of ICFG.mars and DFFG.mcs Interaction.

When an DFFG.mcs is configured to support a given layer 3 multicast group (or set of groups) it interacts with the ICFG.mars in an analogous manner to mclients. Within a given IASG the relationship between the ICFG.mars and its DFFG.mcses is shown in Figure 55~~Figure 5~~. Each DFFG.mcs establishes its own bi-directional point-to-point VC to the ICFG, and uses this to query the ICFG.mars for a group's membership, or register/deregister its willingness to support a given group. This VC may be transient, and represents an instance of the ICFG Control (ICCtl) VC identified in section 4.3.1~~4.3.1~~.

When an DFFG.mcs is supporting a group's traffic it requires rapid updates of group membership changes. To support this the ICFG.mars manages a permanent, unidirectional point-to-multipoint VC, ServerControlVC [*for want of a better name*], out to all DFFG.mcses. ServerControlVC is independent of ClusterControlVC.

Messages on ICCtl and ServerControlVC are LLC/SNAP encapsulated. They use a unique code-point allocated by the IETF to the MARS protocol.

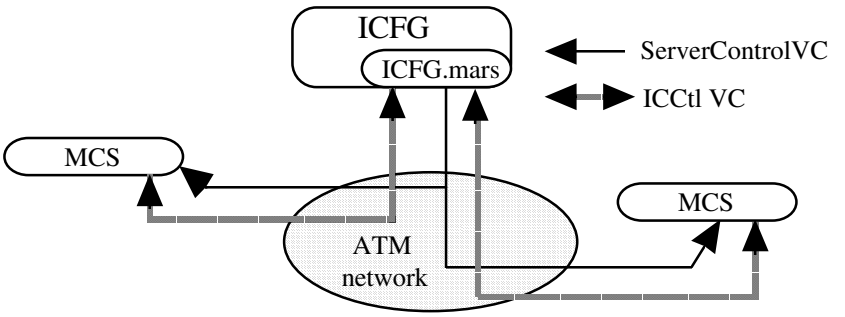


Figure 55 - Control paths between ICFG.mars and MCSs.

The MARS_REQUEST, MARS_MULTI, and MARS_REDIRECT_MAP messages are used by DFFG.mcses in a similar way to mclients.

The following messages are specific to the interactions between an ICFG.mars and DFFG.mcses.

MARS_MSERV,
MARS_UNSERV.
Allow multicast servers to register and deregister themselves with the ICFG.mars.

MARS_SJOIN,
MARS_SLEAVE.
Allow ICFG.mars to pass on group membership changes to multicast servers.

The detailed behavior of the ICFG.mars when handling groups supported by DFFG.mcses is dealt with in a later section.

5.8.4 MPOA Multicast Related LLC/SNAP Code Points.

The LLC/SNAP encapsulation used for ICFG.mars control messages is the same as that used by the IETF's MARS and ATMARP protocols.

```
      [0xAA-AA-03][0x00-00-5E][0x00-03][ICFG.mars control message]
(LLC)       (OUI)      (PID)
```

The extended encapsulation for intra-cluster multicasting of data packets is indicated by another LLC/SNAP codepoint borrowed from the IETF's MARS specification.

```
      [0xAA-AA-03][0x00-00-5E][0x00-01][Extended Layer 3 packet]
          (LLC)       (OUI)      (PID)
```

The extended layer 3 packet is encoded in the following manner:

```
      [pkt$cmi][pkt$pro][Original Layer 3 packet]
```

The first 2 octets (pkt$cmi) carry the Cluster Member ID (CMI) assigned when an endpoint registers with the ICFG.mars. The second 2 octets (pkt$pro) indicate the protocol type of the packet carried in the remainder of the payload. This is taken from the set of protocol identifiers identified in section 5.8.1.15.8.1.1.

When carrying packets belonging to protocols that require the SNAP form of identification the pkt$pro field is extended to carry the 40 bit SNAP field and padding for 32 bit alignment. Such layer 3 packets are encoded in the following form:

[pkt$cmi][0x00-80][mar$pro.snap][padding][Original Layer 3 packet]
2octet 2octet 5 octets 3 octets N octet

For example, an IPv4 packet would be transmitted as:

```
[0xAA-AA-03][0x00-00-5E][0x00-01][pkt$cmi][0x800][IPv4 packet]
```

As direct multicast packet transmission is currently only intra-cluster, no IASG specific identification needs to be carried on a per packet basis. However, future developments may lead to a new extended packet encapsulation format. These will be identified by a different LLC/SNAP codepoint, and may share forward path VCs with data packets encapsulated using the current scheme.

5.8.5 General Control Message Format.

ICFG.mars messages may further be subdivided.

```
[MARS header][Layer 3 and/or ATM addresses][Suppl. Parameters]
```

[MARS header] contains common fields indicating the operation being performed (e.g. MARS_REQUEST) and the layer 3 protocol being referred to.

The format of the following [Layer 3 and/or ATM addresses] area in the MARS message depends on the operation indicated in the [MARS header]. These provide the fundamental information that the registrations, queries, and updates use and operate on.

Finally, [Suppl. Parameters] represents an optional set of supplementary parameters to be associated with the [Layer 3 and/or ATM addresses] information. Examples of supplementary information would be authentication information, QoS 'hints' to associate with a particular mapping, or encapsulation specifications for data VCs created as a consequence of a particular mapping.

Supplementary parameters are intended to act as modifiers of the default behavior(s) associated with the operation indicated in the header. They need not be supplied if the default rules governing the interpretation of the [Layer 3 and/or ATM addresses] information is sufficient at any given time.

The combination of address fields and supplementary parameters allows a range of multicast MPOA Targets and Forwarding Descriptions to be represented.

6. Detailed Component Behaviors

6.1 EDFG Behavior

The EDFG lies between legacy LANs and ATM ELANs as depicted in Figure 66~~Figure 6~~. The EDFG has one or more *LANE interface(s)* and one or more *legacy interface(s)* for each IASG for which the EDFG is configured. Multiple interfaces for a given IASG appear when ELANs for the IASG are bridged together at the edge device.

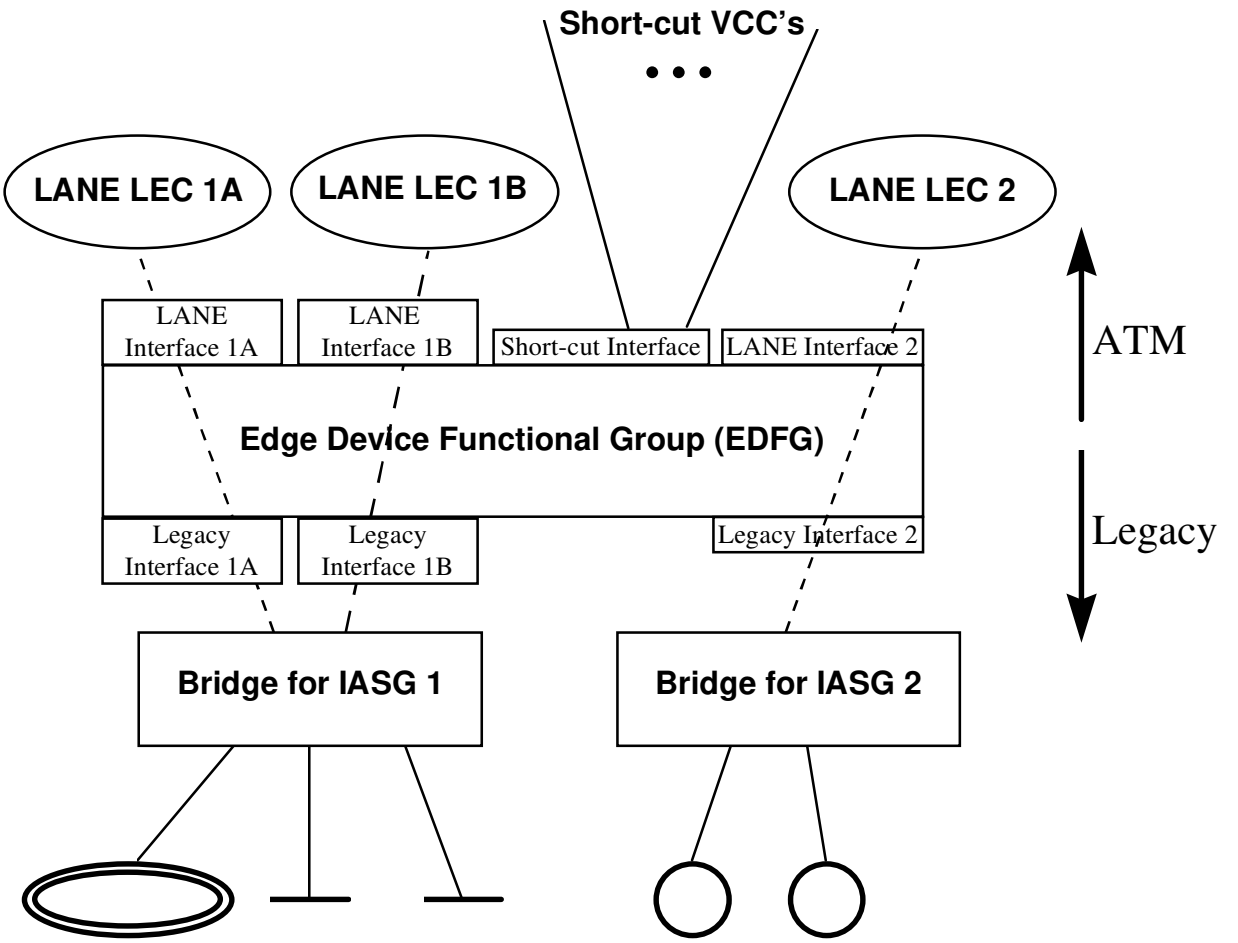


Figure 66 - EDFG Reference Model

The EDFG in Figure 66~~Figure 6~~ intercepts, counts, and/or redirects packets moving between a bridge and the bridge's LANE port (LEC). The difference between a LANE-capable bridge and an MPOA edge device lies in the EDFG. To both the bridge and the LEC, the EDFG is invisible; therefore, both LANE and bridging remain outside the scope of MPOA. The detailed diagram of the EDFG is shown in Figure 77~~Figure 7~~. Note that the EDFG sees only:

1. packets sent by a bridge destined for a LEC (inbound flow); and
2. packets received on a short-cut VCC and relayed to the bridge as if they came from a LEC (outbound flow).

Packets received on a LEC are passed to the bridge without examination by the EDFG.

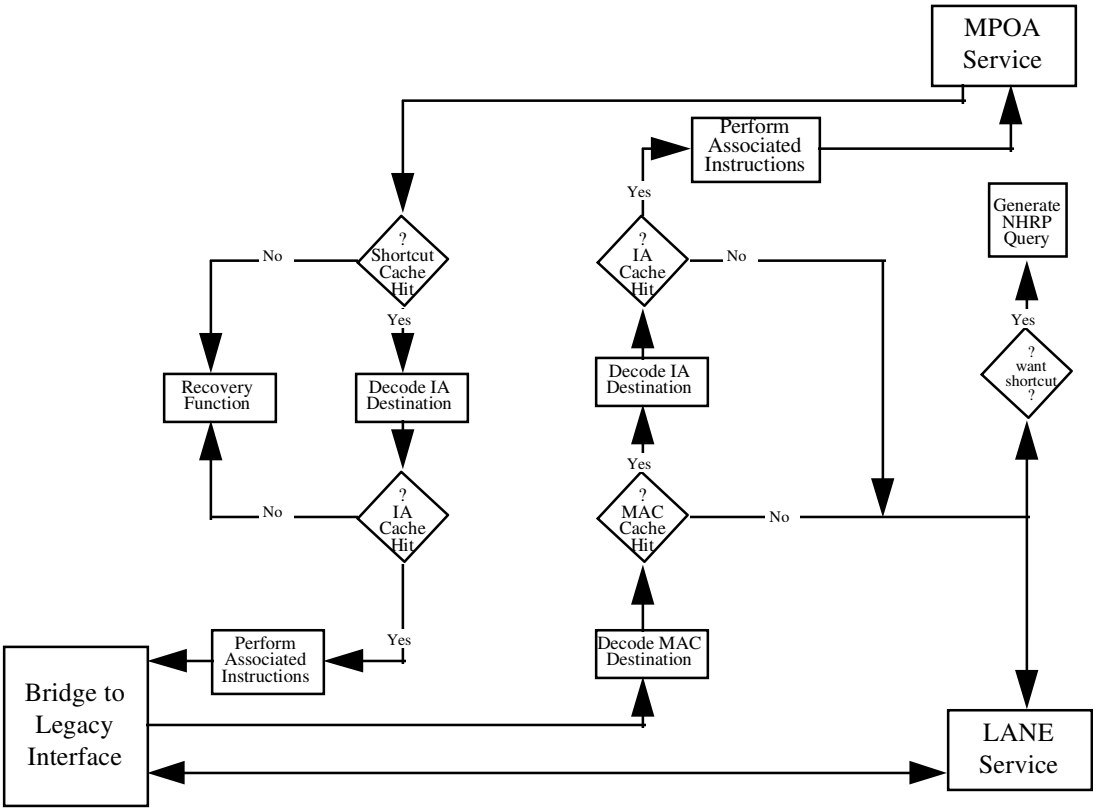


Figure 7~~7~~ - EDFG Behavior Logical Block Diagram

6.1.1 EDFG Inbound Flow

All inbound packets (packets sent by the bridge to a LEC) are examined to see whether they have an “interesting” MAC address. “Interesting” MAC addresses are those MAC addresses which, on the LEC in question, are known from configuration and/or registration information to belong to an RFFG or DFFG. The EDFG examines the internetworking layer destination address of every packet with an “interesting” MAC address, and looks up that {MAC address, internetworking address} pair in its Ingress Cache. The contents of the Ingress Cache are shown in Table 11~~Table 1~~.

Table 1~~1~~: Ingress Cache

Keys		Payload	
MAC address	Internetworking address	count	ATM address or UNI/VCC

If the {MAC address, internetworking address} pair is not found in the ingress cache, a new cache entry is created. The ATM address/UNI/VCC field is invalidated, and the “count” field set to 1 to count the frame. The frame is then sent on to the LEC for output to the ELAN.

If the {MAC address, internetworking address} pair is found, but the ATM address/UNI/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.

When count for a given {MAC address, internetworking address} pair exceeds a configured threshold for number of packets sent within a configured time period, then the EDFG is responsible for sending an MPOA address resolution request to the RSFG or ICFG to which the packet's destination MAC address belongs, requesting a short-cut VCC. [The conditions under which this request may/must be resent are TBD.]

If the {MAC address, internetworking address} pair is found in the ingress cache, and the ATM address/UNI/VCC field specifies an operational VCC, then the packet's MAC header is stripped off, the packet is encapsulated in the appropriate internetworking layer encapsulation, and the packet is sent over the specified short-cut VCC.

6.1.2 EDFG Outbound Flow

For all packets received on a short-cut VCC, the EDFG looks up the UNI/VCC and destination internetworking address in the Egress Cache in order to forward the packet. The contents of the Egress Cache is shown in Table 22~~Table 2~~.

Table 22: Egress Cache

Keys		Payload		
ATM address or UNI/VCC	Internetworking address	IASG-ID	LANE bridge port ID	MAC header

[Editors Note: Key now contains a tag]

If the UNI/VCC and Internetworking address are not in the Egress Cache, the packet is discarded (and the error counted). If the UNI/VCC and Internetworking address are in the cache, but the indicated LANE bridge port is not fully operational (e.g. the bridge is an IEEE802.1D transparent bridge and the LANE port is in the "blocked" state), the packet is discarded (and the error counted). If the UNI/VCC and Internetworking address are in the cache, and the indicated LANE bridge port is fully operational, then the MAC header in the egress cache is attached to the internetworking packet, and the resultant frame is passed to the bridge as if it arrived from the bridge's LANE port.

6.1.3 Legacy to Legacy Flows

An EDFG can support multiple IASGs on its legacy interfaces. As a result, it is quite possible for an EDFG to need to forward a frame between two legacy interfaces on itself at the internetwork level without traversing the ATM fabric. This situation should be true even when the EDFG has multiple ATM interfaces into the ATM network.

If the destination for a frame that has arrived over a legacy interface is another legacy interface on the same EDFG, then the ATM address associated with the destination must represent an ATM address of the EDFG. It is possible for an EDFG to be represented by one or more ATM addresses on the ATM fabric. This is especially true if the EDFG has multiple ATM interfaces.

When the EDFG recognizes one of its own addresses as the destination, for the frame it should then search the cache it is using for ATM to legacy traffic to determine the legacy interface on which to forward the frame. The cache entry in the egress cache that results from the MPOA query needs to be tagged in a manner to identify the source as being on the same EDFG as the target.

It is also permissible, although not recommended, for an EDFG with multiple ATM interfaces to actually establish a shortcut flow across the ATM fabric.

There are several possibilities of how the "turn-around" can occur. The following possibilities are included as illustrative and are not meant to imply any restrictions on the implementation:

- Data can be re-directed directly from the incoming legacy interface to the outgoing interface by logic within the EDFG (this could be drawn as an extension of the logic diagram in Figure 77~~Figure 7~~),
- Data can be redirected by a virtualization abstraction within the box in which the implementor~~implementer~~ "opens a flow" using a call to some function that is able to determine that origin and terminus are at the same address and handles the flow accordingly,
- Data can be redirected by the NIC card which performs a local mapping similar to that used by some switches or
- Data can actually be turned around at the switch - this would be the case if the box does nothing.

6.1.4 Cache Management

[Editors note: Cache Management sections due to be enhanced/modified by the cache imposition work agreed to at the April meeting.]

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.

6.1.4.1 Ingress Cache Entry Management

Ingress cache entries are created from MPOA address resolution responses (solicited or otherwise). When a resolution response is received, the destination internetworking address, terminating ATM address, source holding time, and information relating to encapsulation are used to form the ingress cache entry. If the EDFG originated the resolution request, the ingress cache entry may already exist, but be incomplete.

If the terminating ATM address of the resolution response matches the ATM address of the other end of an existing suitable point-to-point VCC, then the ingress cache entry can be tied to that existing VCC. Otherwise, the EDFG is required to signal the creation of the VCC.

Aging of ingress cache entries uses the source holding time from the latest resolution response received relative to the associated destination internetworking address, but cache entries may be further limited in duration by management/configuration established maximum holding times. In addition, ingress cache entries may be withdrawn by the originating resolution response authority (ICFG or RSFG) at any time.

6.1.4.2 Egress Cache Entry Management

Egress cache entries are created from MPOA address resolution queries. When a resolution authority (ICFG or RSFG) transmits a query to an EDFG, it includes the IASGid and MAC header required to transform an internetworking packet into a MAC frame appropriate to the resolution authority's LEC. The EDFG uses the originating ATM address and destination internetworking address of the resolution query to find and/or create an egress cache entry. If the VCC already exists, then the egress cache entry can be attached to the UNI/VCC, and can be used. If the VCC does not yet exist, then it must be bound to the appropriate egress cache entry (entries) once it is created via signaling. The EDFG receiving the address resolution query is not responsible for signaling the establishment of the short-cut VCC.

Because egress cache entries are used exclusively for delivery of data frames to bridges, they are not affected by either source holding time or withdrawn resolution responses and aging of these entries is a local matter. Note that the encapsulation information associated with any particular egress cache entry and short-cut combination becomes invalid if the short-cut data flow is removed.

An egress edge device may find that it must discard packets received over a short-cut VCC, as the egress cache entry has become invalidated. The details of why a packet must be discarded, or why an edge

device views an egress cache entry as invalidated is a matter local to the edge device. An example would be an edge device which incorporates a bridge running the 802.1D spanning tree protocol, which finds that a packet received over a short-cut VCC is due to be sent over a port which is not in the forwarding state, or is due to be sent back out the LEC port associated with the short-cut VCC that the packet arrived on. This situation could occur due to a topology change. Such a change might result in an edge device no longer being the correct edge device for a given target internetworking address.

If an edge device detects an invalidated egress cache entry it should inform the entity that imposed the egress cache entity in the first place. This could be an ICFG or an RSFG, which will then issue a NHRP Purge Request to the MPOA client that had previously issued the NHRP Resolution Request for the target internetworking address. The MPOA server will delete its own association between the target internetworking address and EDFG ATM address, so that the receipt of a subsequent NHRP Resolution Request will result in the relevant MPOA server finding out again the correct EDFG ATM address to use.

When an MPOA client receives an NHRP Purge Request it must stop using the short-cut VCC for packets destined to the specified target internetworking address. It may issue a new NHRP Resolution Request immediately, or it may wait and use some metrics to determine when to query again.

6.2 ATM Host Behavior

6.2.1 AHFG Behavior

6.2.2 Dual Mode Host Behavior

6.3 ICFG/DFFG Behavior

The ICFG and DFFG together provide the capabilities to coordinate the membership in the IASG and provide communication between AHFG and LANE elements of the IASG. These two functions are always logically co-resident, i.e. information flows between these two are not described in this document. They are described as separate functional groups in order to allow (but not require) them to have separate ATM addresses. This provides flexibility in implementation.

6.3.1 ICFG Behavior

ICFGs receive registration requests of the form (IASGid, ATM address, internetwork address, optional MAC address) from AHFGs. On receiving a request, an ICFG verifies that the requester is talking to the correct ICFG. This checks that the ICFG serves the IASG in the request and that it is the configured ICFG for the AHFG. The ICFG also verifies uniqueness of the information supplied in the request (e.g. MAC address, internetwork address etc.). If these tests succeed, the ICFG stores the information in its tables, and shares the information with all the other ICFGs serving the IASG via the server synchronization protocol (note: this protocol has not yet been defined). The registration request also has an indication whether the AHFG wants the ICFG/DFFG combination to proxy for it on an ELAN. If the AHFG did not provide a MAC address, the ICFG allocates it a MAC address. The ICFG then assigns a working DFFG to the AHFG and returns a success indication and the address of the assigned DFFG to the AHFG. The AHFG will use this DFFG for default data transfers.

Note: The above assumes that the AHFG knows its internetwork address. For AHFGs that acquire their internetwork address from a server (e.g. a DHCP or BOOTP server), some additional mechanisms are needed. These mechanisms are for further investigation.

An ICFG accepts and responds to queries from AHFGs and EDFGs in one of the IASGs it serves for resolving internetwork addresses to ATM addresses. If the target is in the same IASG, the ICFG

responds using local information (from its tables); otherwise, the ICFG forwards the query to an RSFG (the address of the RSFG is provided to the ICFG during configuration). ICFGs also respond to address resolution queries from RSFGs for members of the IASG. When the target is a registered AHFG, the ICFG responds using local information. When the ICFG receives an address resolution query for an as yet "unknown" intra-IASG destination, the ICFG directs its co-resident DFFG to issue appropriate internetwork queries (e.g. IP ARP) in order to get the MAC information. The DFFG will issue the queries to the IASG members using LANE. It is possible that the DFFG will receive multiple responses to these queries (for instance, if it is on a 802.5 LAN; see Annex B for further details). The ICFG processes responses from the target system, and selects the information (e.g. RI field) from one of these responses to use in the MAC header of the frames that it will forward on behalf of MPOA clients to that target. When the DFFG receives the MAC information in response to its queries (e.g. IP ARP Reply), the ICFG resolves the MAC address of the destination to an ATM address using LANE procedures (LE_ARP). Since this ATM address corresponds to a LEC, the ICFG needs further information in order to map the LEC's ATM address to the ATM address of an MPOA component (e.g. an EDFG). Procedures for providing the ICFG with this information are for further study. Once this mapping is accomplished, the ICFG returns the MPOA component's ATM address to the entity from which it received the query.

When an ICFG responds to an address resolution query, the target of the query may reside on a LAN segment behind the EDFG. In this case, the ICFG downloads an egress cache entry (for data frames arriving from the MPOA system and going to legacy LANs) to the EDFG. This cache entry contains at least the (destination internetwork address, MAC address) information that will be needed to encapsulate packets received via a shortcut VCC for transmission on the legacy LAN. ICFGs are responsible for maintaining and updating EDFG cache information. For example, an ICFG may receive messages from RSFGs informing it that cache entries should be invalidated (for instance, due to changes in routing). Ingress cache entries (for data frames arriving from a legacy LAN and going to MPOA) are created by MPOA query responses, whereas egress cache information is created as described above.

An ICFG may also initiate a cache trigger to an EDFG or AHFG in response to DFFG flow detection. In this case, the cache update is, in effect, a command to set up a shortcut VCC.

The ICFG and DFFG also participate in multicast protocols such as IGMP (on behalf of AHFGs). For example, the ICFG/DFFG pair may generate IGMP messages when an AHFG registers for an IP Multicast address. Further details on the extent of the ICFG's role and participation in multicast protocols are required.

6.3.2 Intra-IASG Coordination

Each ICFG is given a list of all other ICFGs supporting a given IASG. It establishes a pt-to-multipoint VC to them, or uses a collection of point-to-point VCs. In either case, whenever an ICFG accepts a registration, it sends the information to all other relevant ICFGs. When a registration is removed (time-out, VC termination, ...) all other relevant ICFGs are notified of that fact as well. Additionally, port status changes (see Spanning Tree below) are also passed on.

When an ICFG gets a query from an AHFG which for a destination not within the IASG, it will pass the query to an RSFG registered as serving the source IASG. (It is for further study to determine whether, if an ICFG is also serving the destination IASG, it can safely deliver the answer without consulting the RSFG.)

editor's note: ICFG coordination will use the IETF SCSP (draft-luciani-rolc-scsp-0x.txt) with enhancements to handle MAC addresses. Text is needed and contributions are requested.

6.3.3 DFFG Behavior

Editor's note: Text needs to be added to describe the following picture and table.

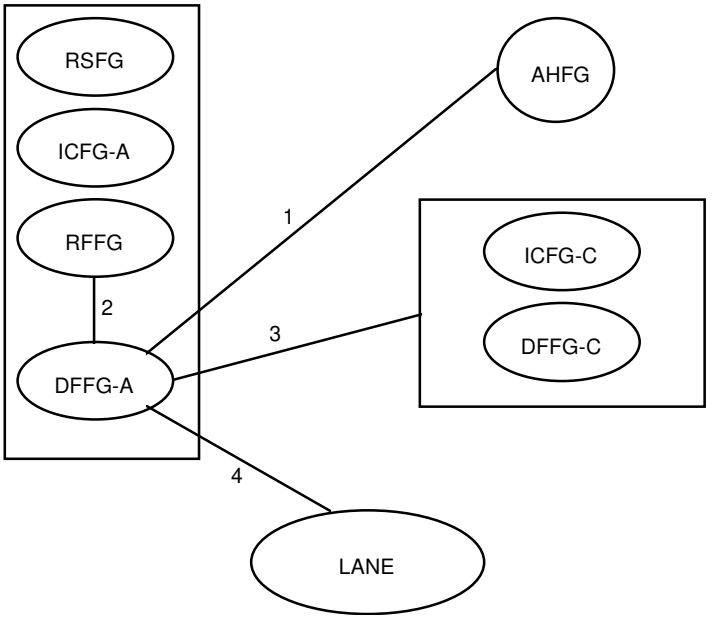


Figure 88 - DFFG Behavior: Model

Table 33: DFFG Behavior

Destination: / Source:	"my" AHFG	intra-IASG AHFG, not "my" AHFG	intra-IASG, not AHFG (LANE)	inter-IASG
"My" AHFG (1)	forward to AHFG	forward to correct AHFG over direct VCC or add MAC and forward to LANE	Add MAC and forward to LANE	forward to RFFG
co-resident RFFG (2)	forward to AHFG	forward to correct AHFG over direct VCC or add MAC and forward to LANE	Add MAC and forward to LANE	discard and count packet
any other non-LANE (3)	forward to AHFG	discard and count packet	discard and count packet	discard and count packet
LANE (4)	strip MAC and forward to AHFG	discard and count packet	discard and count packet	discard and count packet

6.4 RSFG/RFFG Behavior

The RSFG engages in the operation of traditional routing protocols. These may include operation over legacy media reached through EDFG, operation over ATM with identified peers, and operation over ATM Groups using a mesh of point-to-multipoint circuits. The ATM peer operation may be mediated by an IASG/ICFG, or may simply be via configured, numbered or unnumbered, adjacencies. Point-to-multipoint

direct operation would be by configuration, possibly with support from a "multicast client registration function" somewhere. Interaction with legacy media would use LANE.

EDFGs may and ICFGs will send queries to an RSFG. If the destination is in another IASG handled by the RSFG, it will generate the response, and record the fact of the response. If the information is later invalidated, a notification will go to the source of the query. (Due to the absence of a direct VC to an AHFG, the notification to that will go through the support ICFG.) A destination may become invalid either because the actual host moved/expired, or due to a routing change.

If the destination is apparently reachable through another RSFG on the ATM, rather than in a directly served IASG, then the query will be passed to that RSFG. –When the reply is received, state for detecting routing changes is saved, and the reply passed to the original source of the query. In certain scenarios, it may be the case that distributing the information among the RSFGs is an acceptable replacement for the query forwarding mechanism – this requires further study.

If an RSFG receives a query from another RSFG, it checks if it serves the destination IASG or otherwise directs traffic for the destination away from the ATM Fabric. If instead routing points at another RSFG (or indistinguishably an ATM Router) then it passes the query on (and later passes the response back). If it is responding to the query, it saves enough state to notify the relevant AHFG or EDFG if a relevant routing change occurs, and generates the response back to the asking RSFG.

When a routing change occurs, an RSFG checks if it affects any cache management entries. If so, it notifies the source and/or destination EDFG/AHFG that the cache entry for the given destination is invalid. If the “destination” EDFG was notified, and there is still a VC from the source EDFG/AHFG, then a message will go to that source notifying it to invalidate the VC. This is required to avoid routing loops in the presence of routing information. It is insufficient for certain known topologies, and further analysis in this area is needed.

At this time, all answers to queries will be strictly for the requested host. Future extensions to prefix query/response and management will require care in the interaction with routing aggregation, variable subnetting, and routing policies.

6.5 Cache Management

(Editor’s note: This was a topic of conversation and it was agreed that this issue needs to be addressed)

7. Detailed Protocol

[Editor’s note: this section is included for content, but word smithing is required. Once the content of section 7 is stable, we should modify this]

All communication associated with Cache Imposition takes place over the appropriate MPOA Control flow. The identity of each component in such a flow is determined based on the VC over which the flow takes place and the corresponding originating and terminating ATM addresses.

There are two known cases for cache management, egress described in section 6.1.4.2 and ingress described in section 6.1.4.1.

MPOA uses the NHRP (Next Hop Resolution Protocol) defined by the IETF for registration/query/response. This protocol is extended with TLVs to handle the specific requirements of MPOA. Also, MPOA uses the MARS (Multicast Address Resolution Server) defined by the IETF for registering multicast receivers and providing information to multicast senders. The MCS (MultiCast Server) can be used for information replication. This protocol is supported by the ICFG/DFFG for the

services to the AHFGs. As described in the detailed component behaviors, the DFFG is also responsible for the appropriate interaction with the legacy/lane side of the system.

7.1 Egress Cache Management Protocol

7.1.1 Cache Imposition

All MPOA data traffic is in the form of network layer datagrams - either as a result of being de-encapsulated by the ingress device or as a consequence of being originated at a network layer entity (such as an MPOA only ATM attached host). As a result, legacy network traffic must be encapsulated with data link layer information on arriving on a shortcut flow (to an egress EDFG) and prior to forwarding via a legacy interface. Information as to what encapsulation is to be used must be provided to the egress EDFG by the NHS making the NHRP reply on the egress EDFG's behalf.

From the egress EDFG's perspective, it is necessary to be able to associate the layer two encapsulation information thus provided with a particular shortcut flow (which may or may not yet have been established). Therefore, the cache imposition mechanism must provide a means for associating a particular shortcut with a set of encapsulation parameters. This is accomplished by including the ATM address provided to the NHS in the initial NHRP query. This address MUST be the same as will be (or has been) used by the NHRP query originator to establish the shortcut flow. Note that this means the source ATM address must be preserved in NHRP query messages.

An NHS responding to an NHRP request on behalf of an EDFG MUST send an EGRESS_CACHE_IMP message to the EDFG (to provide information needed to establish a cache entry within the EDFG). This would be true regardless of where the NHRP query was originated; however, the NHRP query MUST contain the ATM address the originator will use in establishing the flow.

An NHS waits for an EGRESS_CACHE_ACK prior to returning the initial Next Hop Resolution Response. A cache imposition acknowledge message with a status of success will contain an ATM address to be used in this response. It may also contain an tag TLV to be returned as part of the response as well. If status indicates insufficient resources, the responding NHS may, as a local matter, return some other ATM address (i.e. an ATM address of a local forwarder with an existing flow to the responding MPOA Client) or an error response to the request originator. An NHS need not wait for an acknowledge message on refresh requests.

The egress NHS/RSFG MUST maintain state relative to all valid unexpired cache impositions.

7.1.2 Cache Updates

NHRP relies on refresh as a mechanism for continuing validation of resolution information. Consequently, MPOA clients are required to re-send NHRP Requests on a periodic basis. One result of this is that such requests reaching the egress RSFG trigger an EGRESS_CACHE_IMP message (which SHOULD use the same Cache ID) to update the holding time associated with the corresponding cache entry.

In addition, an RSFG may become aware of a routing change that effects the MAC-layer destination to be used in routing external to ATM (beyond the EDFG egress). In the event that this occurs, the RSFG MUST examine its local state tables for current cache impositions effected by this change and send an EGRESS_CACHE_IMP message to update these impositions. The Cache ID used in these messages MUST be the same as was used in EGRESS_CACHE_IMP message most recently sent to the egress EDFG. In addition, the holding time SHOULD be adjusted to account for elapsed time since the most recent state-update to the NHS/RSFG relative to that cache imposition.

Cache updates, using an EGRESS_CACHE_IMP message with the same Cache ID, must be sent to the egress EDFG using the same VC (or a VC originating from the same NHS/RSFG ATM address and terminating at the same EDFG ATM address) in order to ensure that correct updates are made.

An EDFG receiving an EGRESS_CACHE_IMP message with a Cache ID matching an cache entry previously received from the same RSFG, MUST replace fields in this cache entry with corresponding fields in the new EGRESS_CACHE_IMP message.

7.1.3 Invalidation of Imposed Cache

An EDFG MAY remove any imposed cache entry which has expired un-updated and MUST NOT use the information in such an entry unless and until it is updated with a non-zero holding time by the RSFG that originally imposed it, using an EGRESS_CACHE_IMP message with the same Cache ID. Similarly, an egress EDFG must assume that all cache impositions originating from an RSFG with whom it has lost all communication (all VCs associated with that RSFG have been dropped) and MUST remove these entries.

7.1.4 Invalidation of State Information Relative to Imposed Cache

An NHS/RSFG must assume that all cache impositions made by it to an EDFG with which it has lost all communication are lost and MAY remove state information relative to these impositions. Alternatively, the NHS/RSFG MAY elect to expire this state information normally and re-impose cache associated with remaining state information on restoration of connection to the egress EDFG.

7.1.5 Recovery From Receipt of Invalid Data Packets

An egress EDFG that continues to receive data on a shortcut VC for which it does not have a valid cache entry MUST drop the VC after a finite period of time. This is required in order to recover from situations which may arise as a result of a lost cache imposition or incorrect shortcut usage by the remote end.

7.1.6 Egress Encapsulation

Egress cache impositions will contain DLL encapsulation information when the target is an MPOA EDFG.

7.1.7 EDFG Initiated Egress Cache Purge

The egress cache imposition protocol provides the capability for an EDFG to issue an EGRESS_CACHE_PURGE_REQUEST message, and for a egress cache imposer (ICFG/RSFG) to issue an associated EGRESS_CACHE_PURGE_REPLY message. An egress cache imposer generates an NHRP Purge Request on receipt of an EGRESS_CACHE_PURGE_REQUEST, to the relevant MPOA client. Each EGRESS_CACHE_PURGE_REQUEST message results in one NHRP Purge Request message being generated. If an EDFG has multiple egress cache entries for the same target internetworking address, but with different source ATM addresses, it issues an EGRESS_CACHE_PURGE_REQUEST for each one. When an egress cache imposer receives an associated NHRP Purge Reply it issues an EGRESS_CACHE_PURGE_REPLY message to the relevant EDFG, if such a reply was requested by the EDFG when it sent the request.

Information to be included in an EGRESS_CACHE_PURGE_REQUEST is

- cache ID
- no-reply flag

The cache ID identifies the egress cache entry being invalidated.

The no-reply flag indicates whether or not the EDFG wishes to receive an EGRESS_CACHE_PURGE_REPLY. If set the egress cache imposer must set the "N-bit" in the NHRP

Purge Request, indicating that it does not expect to get an NHRP Purge Reply. If the no-reply flag is cleared the "N-bit" will be cleared, indicating that an NHRP Purge Reply is expected.

Information to be included in an EGRESS_CACHE_PURGE_REPLY is

- cache ID

As before this identifies the egress cache entry.

When an egress cache imposer receives an EGRESS_CACHE_PURGE_REQUEST, it takes the cache id and constructs an NHRP Purge Request. In the intra-IASG case the short-cut is between an AHFG and the egress EDFG. The NHRP Purge Request is, thus, sent to the AHFG that initiated the NHRP Resolution Request.

In the inter-IASG case the short-cut may be between an originating AHFG or EDFG, and the egress EDFG. In this case the NHRP Purge Request is sent to the RSFG (identified by its internetworking address) that re-originated the NHRP Resolution Request after receiving the original NHRP Resolution Request from the MPOA client. The RSFG then forwards the NHRP Purge Request to the originating MPOA client. Note that multiple ingress cache entries may be invalidated as a result of a single egress cache purge request. This is because the scope of the NHRP purge request includes all entries covered by the source, destination and target internetworking addresses in the NHRP purge and is not restricted to the source and destination ATM addresses of the shortcut. Extra TLV extensions may be added to the NHRP protocol to further refine the scope of the purge.

If requested, a NHRP Purge Reply is sent back to the MPOA server that initiated the NHRP Purge Request, which will then generate an EGRESS_CACHE_PURGE_REPLY. If an EDFG does not request an EGRESS_CACHE_PURGE_REPLY it is a local matter to the egress cache imposer whether it requests a NHRP Purge Reply.

7.1.8 Message Contents

EGRESS_CACHE_IMP Message

Information included in a EGRESS_CACHE_IMP message:

- Cache ID
- Destination internetworking address or address prefix
- Ingress ATM address, including
- ATM address type
- ATM address length
- Holding time
- Encapsulation information consisting of LANE DLL header
- ELAN ID
- TLVs

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

Cache ID

Cache ID acts as a transaction ID except where it is desirable to update values sent in an earlier cache imposition message. An EGRESS_CACHE_IMP message may be sent to update an earlier cache imposition by using the same cache ID used in the initial cache imposition operation - indicating which cache entry is to be changed. An MPOA Client receiving such an imposition requires no new resources to be able to acknowledge the imposition. Note that - at present - this method of changing a cache entry seems primarily necessary for an EDFG as it appears to be most useful in updating a cache entry containing a DLL header to be used in forwarding data via a legacy interface after a routing change.

An EGRESS_CACHE_IMP message may also be sent to provide a new holding time as a consequence of a Next Hop Resolution Request sent to refresh a prior Next Hop Resolution Response.

Destination internetworking address/address-prefix

Included in order to provide information needed by the EDFG as a second-stage lookup on legacy bound data-grams. This information is required in order to ensure that the correct encapsulation is identified by the destination internetworking address where datagrams may be received on a flow for more than one such address.

Ingress ATM address

Included in order to allow the MPOA client to associate a subsequent (or existing) VC as the flow used for datagrams destined to the above internetworking address. This information is required in order to setup the first-stage lookup on legacy bound datagrams.

Holding time

The value provided by an EGRESS_CACHE_IMP message should be at least as long as the holding time that will (eventually) be returned in the Next Hop Resolution Response plus some time corresponding to the maximum time expected for the Next Hop Resolution Response to be returned to the originator. Conversely, the MPOA Client may wish to decrement this time value by the amount of time used in processing the cache imposition message. The objective is to ensure that an unrefreshed cache entry becomes invalid some reasonable amount of time after the remote end should have stopped using the shortcut to which it applies.

Encapsulation information consisting of LANE DLL header

The DLL header to be prepended to all datagrams received on the (possibly yet to be established) flow associated with this cache imposition and destined for the above internetworking address. This field, the holding time above and the ELAN ID below may be updated in subsequent EGRESS_CACHE_IMP messages using the same Cache ID. The DLL header should be exactly as it would be if it were to be prepended on a LANE data frame.

ELAN ID

An integer ID number (size yet to be determined) used to differentiate ELANs.

TLVs

In addition Vendor Private Extension TLVs associated with the Next Hop Resolution Request - which were not interpreted by the NHS - may be attached as well. These TLVs may be interpreted by the MPOA Client (if it is able to understand them) and corresponding TLVs of like kind may be returned in an acknowledgment.

EGRESS_CACHE_ACK Message

Information to be returned in a EGRESS_CACHE_ACK is:

- Cache ID
- Status
- Destination ATM address
- TAG TLV

- Additional TLVs

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

Cache ID

This must be copied from the EGRESS_CACHE_IMP message. It is used to correlate imposition and acknowledgment messages.

Status

The following status values may be returned in EGRESS_CACHE_ACK messages:

- 0x00 - Success - cache accepted by this MPOA Client
- 0x01 - Insufficient resources to accept cache
- 0x02 - Insufficient resources to accept shortcut flow
- 0x03 - Insufficient resources to accept either shortcut or cache
- 0x04 - Unspecified/other

Destination ATM address

Determined from the ATM addresses available to the MPOA Client. This information is to be used by the NHS in forming an Next Hop Resolution Response.

TAG TLV

Optionally, an TAG TLV may be returned by the EDFG for use in distinguishing traffic associated with this acknowledgment from all other traffic received on the same flow. This TLV MUST be returned as a Vendor Private Extension with the Next Hop Resolution Response returned by the NHS.

Additional TLVs

Additional TLVs may be attached as well. These TLVs may be interpreted by the NHS (if it is able to understand them) but may not be returned with the Next Hop Resolution Response unless TLVs of like kind were attached to the corresponding request.

7.2 Ingress Cache Management Protocol

MPOA edge devices are required to be able to detect flows. In addition, MPOA servers are permitted to detect flows and request that edge devices establish shortcuts. A trigger mechanism was chosen such that the rest of the protocol will remain consistent with EDFG initiated scenario.

In the event that the need for a shortcut flow has been determined by the MPOA service, the NHS/Router/Route-Server making the determination may trigger the MPOA Client for which it is acting (at present, an RSFG or ICFG may only trigger an EDFG/AHFG that is within an IASG served by it directly) into initiating a Next Hop Resolution Request. This is accomplished using an NHRP_TRIGGER message. The MPOA Client - if it has the resources to establish another short-cut flow (MPOA Clients are required to be able to establish some minimum - non-zero - number of shortcut flows) must respond by initiating NHRP query mechanisms for the MPOA target indicated in the NHRP_TRIGGER message.

7.2.1 Message Contents

NHRP_TRIGGER Message

Information provided in an NHRP_TRIGGER message is:

- Destination internetworking address or address prefix
- MAC address

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

Destination internetworking address/address-prefix

The address to be used in the triggered Next Hop Resolution Request. This address may also be cached for use in second-stage look up (by an EDFG - this address is, in effect, the first/last/only-stage lookup for an AHFG) for legacy originated datagrams which will be sent on the flow to be established as a result of a successful receipt of a corresponding Next Hop Resolution Response.

MAC address

This address is required for an EDFG so that it may perform a first-stage lookup for legacy originated datagrams which will be sent on the flow to be established as a result of a Next Hop Resolution Request. The triggering MPOA service component may provide an invalid MAC address for AHFG clients as these clients MUST ignore this field. In order to provide for indication when a MAC address is "not used" (invalid), "LAN Destination" format as specified in [5] should be used for the MAC address.

7.3 SCSP

The Server Cache Synchronization Protocol (SCSP) is defined in the IETF in draft-luciani-rolc-scsp-0x.txt [WIP]. In order to synchronize its server caches MPOA will use the SCSP mechanism for coordinating the ICFGs and their co-located MARSs within an IASG.

(Editor's note: we need details here and a better title)

8. Signaling Details

(Editor's note: perhaps this is not a separate section, but we need to think about this)

9. References

[1] Grenville Armitage, "Support for Multicast over UNI 3.1 based ATM Networks.", INTERNET DRAFT <draft-ietf-ipatm-ipmc-06.txt>, IP over ATM Working Group, August 1995.

Annex A - Protocol Specific Considerations

For each internetworking protocol mapped onto MPOA, there will be one or more transformations that an EDFG is required to perform between the packet formats used over LANE, and those used over MPOA short-cut VCCs.

Various protocols use different encapsulations using different demultiplexing points (e.g. different LSAPs), and may even have multiple encapsulations on a single MPOA short-cut, the EDFG ingress and egress transformations must be defined on a protocol-specific basis for each internetworking protocol supported.

A1.0 Specific Considerations.

For the IP protocols, there are three formats for IP packets carried over LANE, and one for IP over an MPOA short-cut:

Table 4 - LAN Encapsulations for IP

Type (length)	**Encapsulation** (lengths)								
LANE Ether (16)	LECID (2)	Dest MAC (6)	Source MAC (6)	E'type (2)	**IP packet**				
LANE 802.3 (24)	LECID (2)	Dest MAC (6)	Source MAC (6)	length [10] (2)	LLC (3)	00 00 00 (3)	E'type (2)	**IP packet**	
LANE 802.5 (24-54)	LECID (2)	AC/ FC (2)	Dest MAC (6)	Source MAC (6)	RIF (opt.) (0-30)	LLC (3)	00 00 00 (3)	E'type (2)	**IP packet**
MPOA (8)	LLC (3)	00 00 00 (3)	E'type (2)	**IP packet**					

Given these encapsulations, there is a straightforward form for the MAC header supplied in the egress cache entry: it consists of the entire LANE encapsulation, from LECID through Ethertype. Since there is only one encapsulation for IP on an MPOA shortcut, the transformations performed are not difficult. The MPOA encapsulation is traded with the complete LANE encapsulation.

The biggest difficulty lies in the LANE 802.3 format. This MAC header requires a length field indicating the length of all fields, starting with the LLC field, and including the IP packet. Therefore, the egress cache MAC/encapsulation header is not transparent to the EDFG. The EDFG must parse the egress cache header at least once, to determine that the length field is present, and must insert the correct value for the

[10] Per the LANE specification, this length is set to zero (0) whenever the length of the frame (starting from the LLC header and not including the AAL5 pad and trailer) is 1536 bytes or greater. Otherwise it contains the length of the frame.

length of each outbound packet. Similarly, for inbound packets, the EDFG must parse the LANE frame to find the IP packet and check the validity of the length field), and then perform the simple transformation to the MPOA format.

Annex B - Source-Routing Support in an MPOA Service

B1. Overview of Source Route Bridging

The general format of a token-ring frame is shown in Figure 99~~Figure 9~~. A major difference between LANs that support source-routing such as token-ring and other LAN types is the optional inclusion of the routing information (RI) field (up to 32 bytes) after the destination and source addresses

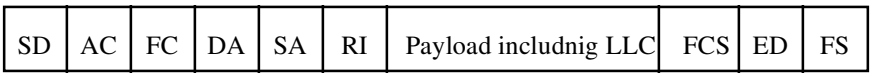


Figure 99 - Token Ring Frame Format

The RI field defines the path a frame will take through the interconnect network of bridges and ring segments. The RI field is built by the end system originating the frame and it is processed by the bridges attached the network to determine if and how the frame should be forwarded. The first two bytes of the RI field is a control field that indicates the following:

- type of explorer frame - all routes or spanning tree only
- length of the RI field
- direction in which the RI field is to be processed
- largest frame that can be sent along the path defined by the RI field

The remainder of the RI field, up to 30 bytes, is composed of 2 byte route designator (RD) fields. The RD field consists of a ring number portion and a bridge number portion. The resulting sequence of ring numbers and bridge numbers define a unique path through the bridged network. The bridges in a SRB network examine the RD field to determine if the frame should be forwarded by the bridge. In the case of explorer frames, a bridge adds its bridge number and the next ring segment to the RI field and increases the length of the RI field as the frame traverses the network.

A bridge or other device supporting source route bridging must support three modes of forwarding frames: specifically routed frames (SRF), spanning tree explorer (STE) frames and all routes explorer (ARE) frames. Specifically routed frames contain a complete RI field that the bridge must examine to determine if it must forward the frame. Explorer frames are used by end systems to discover and build the RI field. Spanning tree explorer frames, as the name implies, are only forwarded by those bridges that are part of the spanning tree created by the bridges. All routes explorer frames are forwarded by all bridges in the network unless the frame has already traversed the next ring segment in the network. All routes explorer frames allow an end system to discover if multiple paths exist to the destination. The end system can then choose the path most suited to its needs: frame size, delay, number of hops, etc.

An important characteristic of SRB networks is that the presence of the RI field allows multiple end systems with the same MAC address to exist in the same bridged network. The only restrictions are that duplicate MAC addresses reside on separate ring segments and that the end systems with the duplicate MAC addresses provide the same service.

B2. SRB Considerations

In order to provide support for source route bridging, some additional function needs to be added to the ICFG/DFFG components. Specifically, a mechanism is needed to obtain source routing information for AHFGs. The RI field is used to define a specific path through the bridged network. Additionally, for non-routable protocols, the RI field is part of the "network" header that should be included in all frames. Also, a

mechanism is needed to respond to explorer frames, and receive specifically routed frames on behalf of AHFGs that are not capable of doing so themselves.

B3. Source Route Bridging Support in MPOA

An ATM-attached host with both MPOA and LANE stacks (i.e., with an AHFG and an LEC) may simply use its LEC to leverage LANE functions defined for source routing (note: the same applies to an ATM-attached host with only a LANE stack; however, this is outside the scope of MPOA). However, in order for the host's AHFG to participate in functions that involve source routing, MPOA procedures are necessary. The same procedures are applicable when an ATM-attached host with only an AHFG requires source routing support. Additionally, an EDFG must support cut through connections to/from AHFGs and other EDFGs on which it is receiving and sending frames in the internetwork format. In the following discussion, we identify the situations in which source routing support is required, and outline the procedures for providing this support.

B3.1. AHFG and ICFG/DFFG Behavior

When an AHFG needs to send data to a device on a source route bridged network, a mechanism is needed to obtain the RI field to include in the frames' headers. This mechanism must be part of the MPOA address resolution process. The following actions occur when an AHFG has a frame to send:

- When an AHFG wishes to send traffic to an internetwork destination, the AHFG may directly send to the DFFG the packet that it wishes to transfer to the internetwork destination. It is the ICFG/Duff's responsibility to perform address resolution procedures on behalf of the AHFG in order to obtain the MAC header to prepend to the packet. Alternatively, the AHFG may send an MPOA address resolution request to the ICFG in order to determine the ATM address to use. (Note: if it is determined that the internetwork destination is outside the IASG, the AHFG may send its query directly to the RSFG. However, in that case, source routing support is not required.)
- If the ICFG is able to locate the destination address in its (MPOA) tables, the ATM address of the destination is returned to the AHFG. If the ICFG is unable to find the destination internetwork address in its (MPOA) tables, and the destination is within the IASG, the MPOA address resolution request must be converted to a internetwork protocol address resolution packet (e.g. IP ARP), and broadcast to all members in the IASG. The ICFG/DFFG does this by broadcasting the packet on all the ELANs in the IASG using LANE procedures.
- If it is on a 802.5 LAN, the ICFG/DFFG may receive multiple responses to this query.
- The ICFG/DFFG processes all responses from the target system, and selects the RI field from one of these responses to use in the MAC header of the frames that it will forward on behalf of the AHFG.
- If the protocol supported by the IASG is a non-routable protocol that has been promoted to internetwork status, then the RI field is returned to the AHFG for inclusion in frames as part of the "internetwork address of the protocol". If the AHFG has sent an MPOA address resolution request to the ICFG, then the ICFG returns the ATM address of the ICFG/DFFG to the AHFG. The AHFG establishes a VC to the ICFG/DFFG and begins sending data. The ICFG/DFFG will forward frames from the AHFG to the target after including the MAC header in them. The ICFG/DFFG will use LANE to forward these frames.

The ICFG/DFFG is also responsible for responding to explorer frames for the AHFGs for which it is serving as a proxy on an ELAN.

B3.2 EDFG Behavior

The EDFG behavior in support of SRB networks must be in accordance with the IEEE 802.1 specification [C2] and the LAN Emulation specification [C1]. This includes assignment of bridge numbers and ring numbers, and processing of explorer frames and specifically routed frames. This behavior does not need to be addressed by MPOA.

Additionally, an EDFG must support cut through connections to/from AHFGs and other EDFGs on which it is receiving and sending frames in the internetwork format. When the legacy network is supporting source routing, the EDFG is responsible for maintaining an RI field that is included as part of the MAC header sent on the legacy network.

On frames received from the legacy networks, the EDFG strips the MAC header including the RI field, unless the RI field is part of the internetwork address, before sending the frame on the MPOA direct connection.

Editor's note: this should eventually be folded into the main text.

B4. References

[C1] Ellington, B. (editor), "LAN Emulation Over ATM: Version 1.0 Specification," ATMF/94-0035R9, January 1995.

[C2] ISO/IEC 10038; ANSI/IEEE Std. 802.1d, "Information Processing Systems - Local Area Networks - (MAC Bridges).

Annex C - NHRP Assumptions

1. Hold down timers must be present in and apply to resolution responses and other protocol messages which may be used to construct cache entries.
2. At least one ICFG s required to be co-resident with each RSFG for each IASG served by that RSFG. (Means that don't have to have an ICFG to RSFG protocol).
3. Any resolution request which needs to be forwarded and which originated at an EDFG must be stashed and re-originated by the RSFG serving the request.
4. All resolution responses are "authoritative"
    - however, not all resolutions responses are "stable
5. Caching of resolution responses at intermediate systems is not allowed.
6. MPOA resolution protocol must provide shortcuts within a single routing domain.
7. MPOA resolution protocol may provide shortcuts within multiple routing domains under certain circumstances (TBD).

(Note: This section needs more text and explanation or needs to be moved into the body of the document)

Appendix A - Issues Surrounding MPOA Support for Layer 3 Multicasting

A1. Multicast Issues Introduction

In addition to unicast, MPOA will have to address the need for multicast and broadcast services by layer 3 data protocols. This contribution is intended as an introduction to the sorts of issue we will face in this area. Specification of actual solutions will occur in further contributions later this year. The ideas in this contribution owe a lot to the discussions held on the IETF's IP-ATM working group mailing list over the last 10 months. An overview is provided of this related work in the IETF', along with pointers on how to access their latest documents on multicast over ATM.

The intention of the MPOA working group is to provide multicast information by referencing the IETF's multicast RFC when it is completed. If the working group determines that this RFC is insufficient to meet the needs of MPOA, the group will raise the issues at the IETF and pursue a joint solution. In the case where differences remain, this document will reference the RFC and point out differences (deltas).

The bulk of the IETF's work is conducted on mailing lists. Multicast specific work is conducted on IP over ATM working group's list. To added to this list, send mail to ip-atm-request@matmos.hpl.hp.com. To send mail to the entire list, send mail to ip-atm@matmos.hpl.hp.com.

A1.1 Some Definitions.

Unicast, multicast, and broadcast may be defined in terms of the number of destinations reached by a single transmit operation of a single PDU by a source interface. In this contribution they are defined as follows:

Unicasting - the PDU reaches a single destination.

Multicasting - the PDU reaches a group of one or more destinations.

Broadcasting - the PDU reaches all destinations.

The astute will notice that both unicast and broadcast transmissions are special cases of multicast, with a groups of one or all destinations respectively.

(In practice the definition of broadcast is supplemented by some statement defining the scope of 'all destinations'. In LAN environments the physical size of a link level segment usually defines the scope of a broadcast. At the layer 3 level (e.g. IP) the scope of a broadcast is usually defined by the scope of a unicast subnet (however 'subnet' is defined)).

We are aware that with ATM networks it can be foolish to define 'all destinations' to cover every attached ATM interface. In practice the interfaces attached to an ATM network are usually already divided into logical or virtual layer 3 (sub)networks for unicast purposes, providing a reasonable basis on which to define the scope of a broadcast.

For the purpose of this contribution an abstract term will be borrowed from the current IETF work in this area [A1]:

Cluster - the set of ATM interfaces that are willing and able to participate in direct ATM connections to achieve multicasting and broadcasting of AAL_SDUs between themselves.

A1.2 The General Problem for MPOA.

A number of questions require answering before we move towards specifying a solution:

- How do we achieve efficient multicasting of AAL_SDUs around a Cluster (intra-cluster)?
- How do the solutions we look for here interact with our unicast solutions? (and should the unicast solution simply be a subset of the general multicast solution?)
- What minimal UNI functionality will MPOA expect to work with?

The last question has a large impact on the complexity of MPOA clients. However, we do not really have many choices. If we wish MPOA to have a solution useful to the present and growing base of installed ATM networks, we must support intra-cluster multicasting using UNI 3.1.

The second last question has an obvious technical answer - ~~–~~that we should build the MPOA unicast and broadcast support around a core, general multicast mechanism. However, it poses an equally obvious practical and political problem - most people are already further down the development path for unicast-only solutions.

A2. Intra-cluster multicasting using UNI 3.1

Two key areas need to be addressed - the data path, and group management.

The data path covers the actual mechanism(s) used to achieve the replication and redistribution of a source's AAL_SDUs to the destinations. Group management covers the mechanisms used to ensure the set of destinations can change over time, in response to the changing needs of the overlying Layer 3 protocol(s).

A2.1 Data path - methods of multicasting.

Given the definition of multicasting in section 1.1, the source must be able to execute a single transmit operation to multicast an AAL_SDU. Assume the source is an AAL User passing an AAL_SDU to a single instance of an AAL5 service. This implies that traffic exits the source interface and crosses the UNI on a single VC - multicasting occurs beyond the UNI.

UNI 3.1 provides the ability for a source interface to originate a point-to-multipoint VC, which is exactly what our model requires from the source's perspective. Using this basic service, two key models may be implemented - 'multicast VC meshes' and 'multicast servers'. In both cases the AAL_SDUs are passed transparently from source to destination(s), without interpretation of their contents at any point along their path.

The most fundamental approach, given the existence of point-to-multipoint VCs under UNI 3.1, is the multicast mesh, shown in Figure 1010~~Figure 10~~(a). This approach results in a data path from source to destination(s) where the actual multicasting occurs down at the cell level, within the ATM cloud's switch fabrics. Multipoint-to-multipoint communication is achieved when multiple senders establish their own data paths to the same set of leaf (receiving) nodes, giving rise to the name 'mesh'.

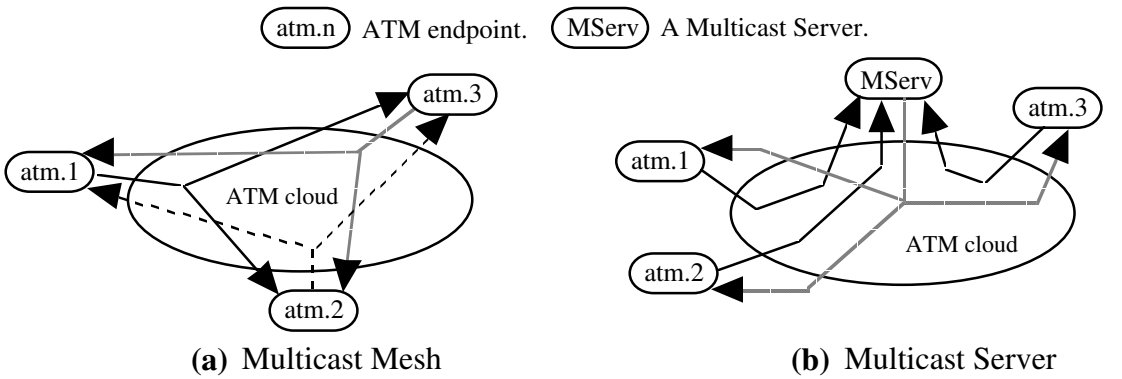


Figure 10~~10~~ - ATM Level Multicasting - Distributed Meshes or Central Servers

The 'multicast server' model is shown in Figure 10~~Figure 10~~(b). In this model the source(s) establish a point-to-multipoint VC with only a single leaf - the multicast server. The multicast server itself establishes and manages a point-to-multipoint VC out to the actual desired destinations. (Although a point-to-multipoint VC is not strictly necessary in Figure 10~~Figure 10~~(b), more complex architectures may use more than one multicast server to distribute the traffic for a given layer 3 multicast group. In this case each multicast server would be one leaf node on the source's outgoing VC).

AAL_SDUs sent to the multicast server are reassembled and then retransmitted back out to all endpoints that have been identified as receivers of this traffic. The leaf nodes of the multicast server's point-to-multipoint VC must be established prior to packet transmission, and the multicast server requires an external mechanism to identify them. (An alternative method is for the multicast server to explicitly retransmit packets on individual VCs between itself and group members. A benefit of this second approach is that the multicast server can ensure that sources do not receive copies of their own packets. However, this is significantly less efficient.)

Reassembly of incoming AAL_SDUs is required at a multicast server as AAL5 does not support cell level multiplexing of different AAL_SDUs on outgoing VCs.

A2.2 Group management and addressing.

Under UNI 3.1 a point-to-multipoint VC is established and managed by the root (source) interface. Only the root may add or drop leaf nodes on this VC, and it must explicitly identify the leaf nodes using their unicast ATM addresses. There is no concept of an ATM group address under UNI 3.1.

This is the core of an essential mismatch between UNI 3.1 and the requirements of protocols such as IP. The IP multicasting model is receiver controlled - leaf nodes decide when they'll become leaf nodes [A2]. UNI 3.1's model is source controlled - the data path only changes when the root issues an ADD_PARTY or DROP_PARTY.

This mismatch imposes a requirement that MPOA's solution has the ability to propagate the group join/leave intentions of Layer 3 entities back to the active source(s) for the group being joined or left. Further, this propagation mechanism must be timely.

Layer 3 data protocols that support the notion of multicasting usually do so through the notion of 'group addresses' - a destination address that indirectly identifies the set of layer 3 group members to which a packet must be sent. A node becomes a source as soon as it is required to transmit a packet with a layer 3 group address as its destination. At this point the source will need to ascertain what unicast ATM addresses correspond to the leaf nodes of the group identified by the specified layer 3 destination group address.

The MPOA solution must therefore provide a mechanism for a new source to find out who all a group's current leaf nodes are.

The MPOA multicast solution will need to support the mapping of arbitrary layer 3 addresses to sets of one or more {ATM.1, ATM.2, ..., ATM.n} addresses. Further, the addresses identified by ATM.n may specify either an NSAPA, E.164, or combined Called Party Number and Called Party Subaddress.

A2.3 VC meshes and Multicast Servers - a time and place for either.

Arguments over the relative merits of VC meshes and multicast servers have raged for some time. Ultimately the choice depends on the relative trade-offs a system architect must make between throughput, latency, congestion, and resource consumption. Even criteria such as latency can mean different things to different people - is it end to end packet time, or the time it takes for a group to settle after a membership change? Ultimately the decision depends on the characteristics of the applications generating the multicast traffic.

If we focused on the data path we might prefer the VC mesh because it lacks the obvious single congestion point of a multicast server. Throughput is likely to be higher, and end to end latency lower, because the mesh lacks the intermediate AAL_SDU reassembly that must occur in multicast servers. The underlying ATM signaling system also has greater opportunity to ensure optimal branching points at switches along the multicast trees coming out of each source.

However, resource consumption will be higher. Every group member's ATM interface must terminate a VC per source (consuming space for state information, instance of an AAL service, and buffering in accordance with the vendors particular architecture). On the contrary, with a multicast server only 2 VCs (one out, one in) are required, independent of the number of senders. The allocation of VC related resources is also lower within the ATM cloud when using a multicast server. These points may be considered to have merit in environments where VCs across the UNI or within the ATM cloud are valuable (e.g. the ATM provider charges on a per VC basis), or AAL contexts are limited in the ATM interfaces of endpoints.

If we focus on the signaling load then multicast servers have the advantage when faced with dynamic sets of receivers. Every time the membership of a multicast group changes (a leaf node needs to be added or dropped), only a single point-to-multipoint VC needs to be modified when using a multicast server. This maps to a single ADD_PARTY or DROP_PARTY message across the UNI. When your multicast service is based on a mesh, an ADD or DROP action is triggered at the UNIs of every traffic source - the transient signaling load scales with the number of sources. This has obvious ramifications if you define latency as the time for a group to stabilize after change.

Finally, multicast servers introduce a 'reflected packet' problem. Sources that are also group members will get copies of their own packets straight back from the multicast server. The MPOA solution must ensure sufficient information is carried within each PDU to enable identification and removal of these reflected packets (this problem was faced and addressed in LANE by the use of a LEC_ID in the packet encapsulation - a more general, LLC/SNAP based solution, is being discussed within the IP-ATM working group [A3]).

The MPOA solution should allow system administrators to utilize either approach on a group by group basis.

A2.4 Broadcast as a special case of multicast.

Treat a layer 3 specific 'broadcast' address as referring to a group containing the entire membership of a Cluster. Allow the MPOA multicast address mapping service to return all members of a Cluster in response to a query for the members of the Layer 3 protocol's 'broadcast' group.

Layer 3 broadcasts are dangerous things to encourage. However, the transition of existing legacy applications will be expedited if support for this is included in MPOA. It is trivial to include it once the general multicast support exists.

A3. Summary of Demands Placed on MPOA's Multicast Solution.

- If we wish MPOA to have a solution useful to the present and growing base of installed ATM networks, we must support intra-cluster multicasting using UNI 3.1.
- We should consider building the MPOA broadcast support around a core, general multicast mechanism.
- The MPOA multicast solution will need to support the mapping of arbitrary layer 3 addresses to sets of one or more {ATM.1, ATM.2, ..., ATM.n} addresses.
- The MPOA multicast solution requires the ability to propagate the group join/leave intentions of Layer 3 entities back to the active source(s) for the group being joined or left.
- The MPOA multicast solution should allow system administrators to utilize either multicast servers or VC meshes on a group by group basis.
- The MPOA packet encapsulation mechanism must provide support for the detection and filtering of reflected packets when multicast servers are in use.

A4. Appendix A References

[A1] Grenville Armitage, “support for Multicast over UNI 3.1 based ATM Networks.”, INTERNET-DRAFT <draft-ietf-ipatm-ipmc-04.txt>, IP over ATM Working Group, February, 1995.

[A2] S. Deering, “Host Extensions for IP Multicasting”, RFC 1112, Stanford University, August 1989.

[A3] G. Armitage, “Issues surrounding a new encapsulation for IP over –ATM”, INTERNET-DRAFT <draft-armitage-ipatm-encaps-00.txt>, IP over ATM Working Group, April 1995.

[A4] T. J. Smith, G. Armitage, “IP Broadcast over ATM networks”, INTERNET-DRAFT <draft-smith-ipatm-bcast-00.txt>, IP over ATM Working Group, April, 1995

[A5] G. Armitage, “Using the MARS to support IP unicast over ATM”, INTERNET-DRAFT <draft-armitage-ipatm-mars-unicast-00.txt>, IP over ATM Working Group, April, 1995.

For Further Reading:
http://gump.bellcore.com:8000/ ~gja/i-drafts.html

Appendix B - Examples of MPOA Control and Data Flows

B1.0 Introduction

The purpose of this appendix is to describe a set of flows between different pairs of MPOA end-systems. Each flow is described for the *default path*, the path which does not resort to an MPOA shortcut, as well as the *shortcut path* which is established following completion of a Next Hop Resolution Protocol (NHRP) address resolution and cache imposition.

B1.1 Data Discard Due to Port Blocking

The MPOA edge device (ED) includes one or more bridging entities performing IEEE 802.1d bridging functionality. This bridging component may cause data to be discarded due to 802.1d blocking of a port. This may occur in the inbound direction if an ATM port is blocked as well as in the outbound direction if a legacy port is blocked. This functionality is mentioned briefly here rather than explicitly for each scenario of this appendix.

B1.2 NHRP Purge Due to Port Blocking

It is possible for an existing shortcut to become obsolete due to a bridging topology change. When an edge device receives outbound data from a shortcut which is destined to a blocked legacy port, it generates an NHRP Purge. The purge is sent to the RSFG serving the egress edge device. The RSFG ensures that the purge is passed to the MPOA client that originated the packet causing that client to remove its cache entry for the shortcut. Throttling of purges is TBD.

B2.0 Scenarios

A multitude of scenarios are considered in this section to describe an almost exhaustive list of MPOA/LANE flows. The simple MPOA network configuration shown in Figure 11 is used as a basis for description of the flows. The MPOA network consists of two IASGs, namely, IASG-1 and IASG-2. The IASG-1 consists of two emulated LANs (ELANs), namely ELAN-1 and ELAN-2. Note that the two ELANs are added to introduce a little more complexity! The IASG-2 contains a single ELAN. Both IASGs contain one or more MPOA end-systems. The MPOA end-systems are classified as AHFG, Dual-Stack Host, and Edge Device. A Dual-Stack Host is assumed to have a resident LAN emulation client (LEC) and an AHFG. Attached to the legacy ports of the edge devices are hosts, labeled from H1 to H7. The IASG-1 and -2 are supported by separate ICFG/DFFG, both of which connect to a single RSFG/RFFG. Note that, the ICFG 1/DFFG 1 has two proxy-LECs, one participating in ELAN-1 and the other in ELAN-2.

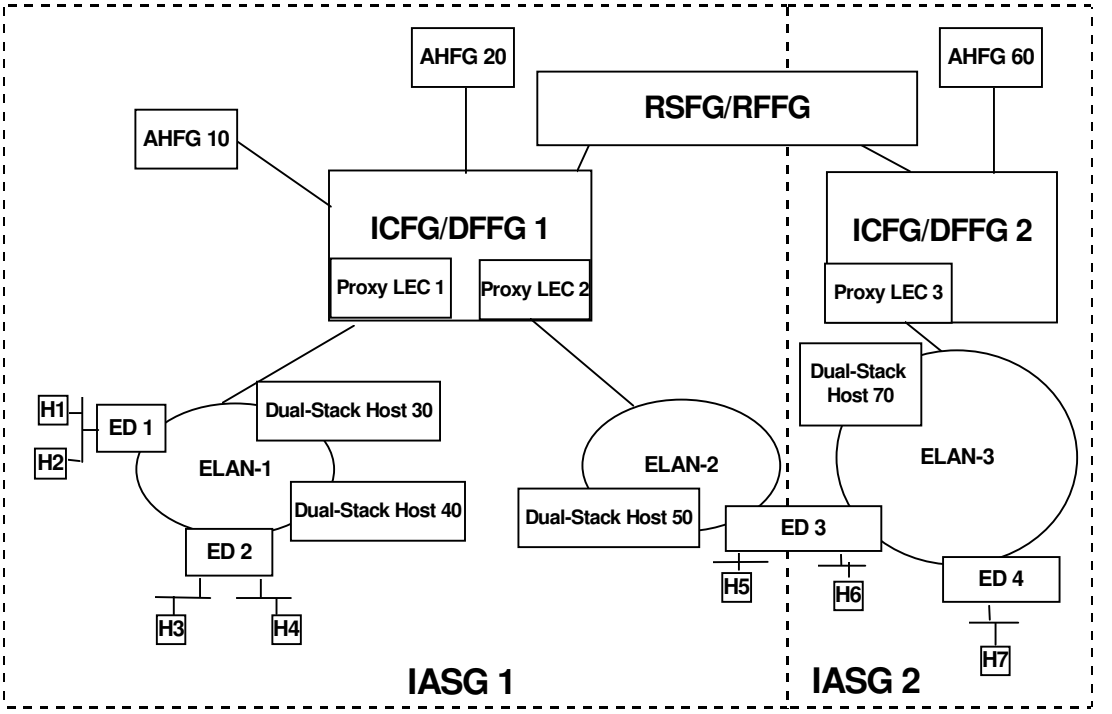


Figure 11 - Example Network Configuration

In order to describe each flow, a pair of end-systems are chosen from Figure 11. The pair of end-systems are chosen within the same IASG or in different IASGs. Thus, the flows are grouped as the intra- and inter-IASG flows. The intra-IASG flows are further sub-classified as intra-ELAN and inter-ELAN flows to account for the fact that there could be multiple ELANs within an IASG.

B2.1 Assumptions

The following assumptions are made in the diagrams although they are optional:

1. NHRP_TRIGGER by the ICFG to the Edge Device.
2. Shortcut initiated by Dual Host to a destination.
3. ***Cache_Imp*** and ***Cache_Ack.***
4. A bridge is indicated for Inter-ELAN forwarding, the location of this is currently outside MPOA specification.
5. *AHFGs may perform mask and match.*

B2.2 Intra-IASG Scenarios

The Intra-IASG flows originate from an AHFG, edge device or a dual-stack host. Similarly, the flow terminates on an AHFG, edge device and the dual-stack host. The matrix shown in Table 5 illustrates all source-destination pairs with the matrix entry representing the scenario-index.

B2.2.1 Intra-ELAN

Table 5 assumes that the flows are within the same IASG. Also, both the dual-stack host and the edge device have a port on the same emulated LAN. Note that the source and destination are different end-systems. Thus, the trivial scenario of legacy-legacy flow on the same edge device and within the same ELAN is not covered.

Table 5: Intra-IASG Scenarios

destination → / source ↓	AHFG	Dual-Stack Host	Edge Device
AHFG	(a)	(b)	(f)
Dual-Stack Host	(c)	(d)	(m)
Edge Device	(g)	(l)	(j)

B2.2.2 Inter-ELAN

Table 6 covers additional flows assuming that the source-destination pair are in different ELANs. Please note that the AHFG row and column are shaded as the ELAN choice does not apply to the AHFG.

Table 6: Inter-ELAN Scenarios

destination → / source ↓	AHFG	Dual-Stack Host	Edge Device
AHFG			
Dual-Stack Host		(e)	(p)
Edge Device		(n)	(k)

B2.3 Inter-IASG Scenarios

The flows listed in Table 7 are between the source-destination pairs for which the source and destination are in different IASGs.

Table 7: Inter-IASG Scenarios

destination → / source ↓	AHFG	Dual-Stack Host	Edge Device
AHFG	(A)	(B)	(E)
Dual-Stack Host	(C)	(D)	(P)
Edge Device	(F)	(N)	(G, H)

B3.0 Flows

B3.1 Intra-IASG

B3.1.1 AHFG Flows

The intra-IASG flows are covered starting with AHFG flows to other MPOA end-systems as illustrated in Figure 12.

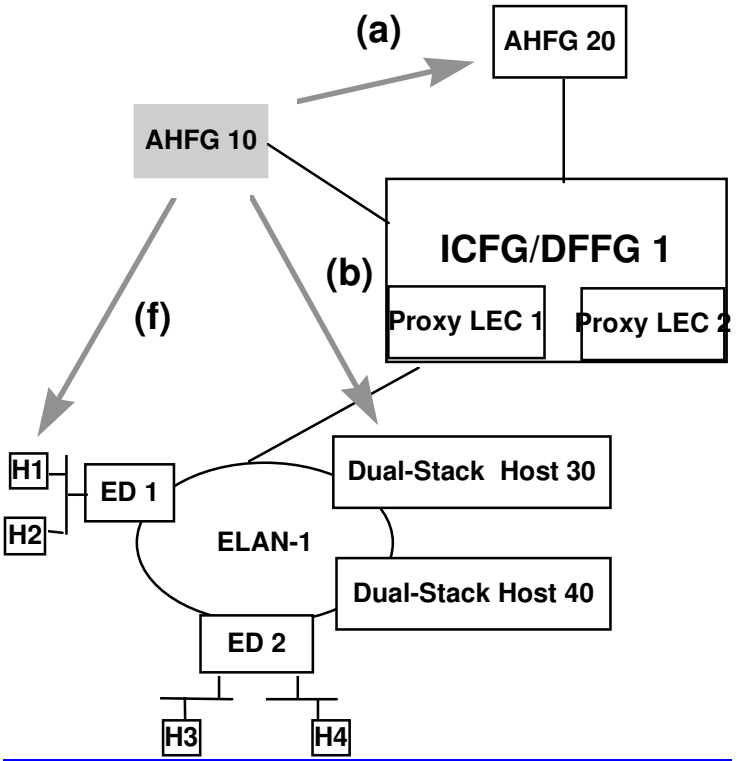


Figure 12 - Intra-IASG AHFG Flows

B3.1.1.1 Scenario (a): AHFG-10 to AHFG-20

For data originating from AHFG-10 and destined to AHFG-20 within the same IASG, Figure 13 shows the default and shortcut paths.

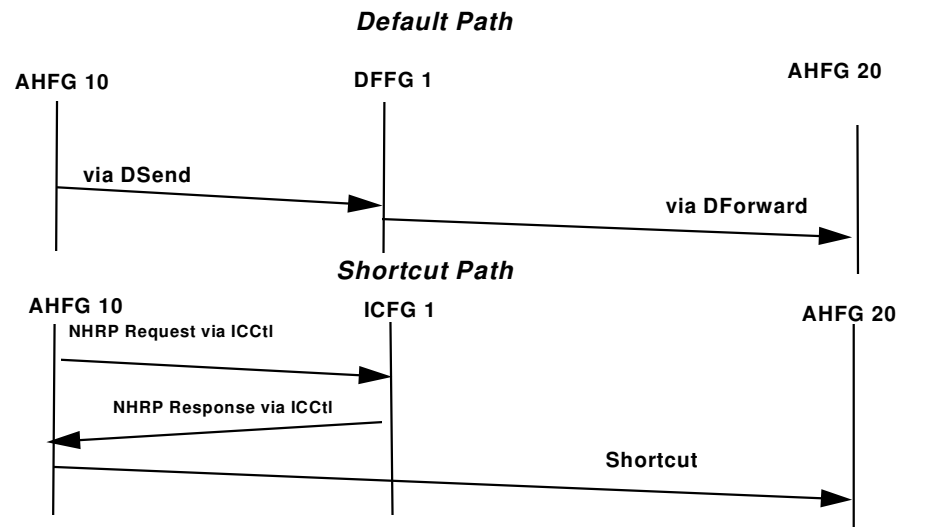


Figure 13 - Scenario (a)

Default Path:

AHFG-10 either does not have a cache entry or does not have a shortcut for the destination internetworking address of AHFG-20. If no entry exists for the destination internetworking address, an entry is created for the purpose of flow detection. AHFG-10 performs an increment on the cache entry's

"count" field and flow detection processing is performed. If a flow is not detected, the internetworking protocol packet is then passed to the DFFG with which the AHFG registered (DFFG-1) via DSend. DFFG-1 then sends the internetworking protocol packet to AHFG-20 via the DForward path.

Shortcut Path:

If AHFG-10 detects a flow to the internetworking address of AHFG-20, it sends an NHRP Request to ICFG-1 via ICCtl. ICFG-1 then sends an NHRP Response to AHFG-10 via ICCtl. AHFG-10 may then establish a shortcut VC to AHFG-20 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to AHFG-20, AHFG-10 encapsulates the internetworking protocol packet with the appropriate encapsulation for the shortcut. The packet is then sent to the destination AHFG using the VC specified in the Inbound Cache entry.

B3.1.1.2 Scenario (b): AHFG-10 to Dual-Stack Host-30

The flow for this scenario is similar to scenario (a).

B3.1.1.3 Scenario (f): AHFG-10 to Legacy Host H1

For data originating from AHFG-10 and destined to Legacy Host H1 within the same IASG, Figure 14 shows the default and shortcut paths.

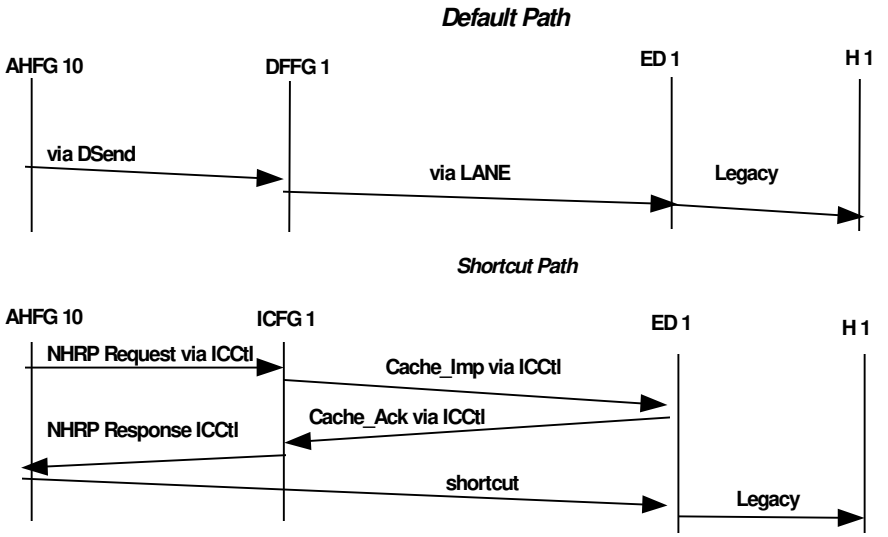


Figure 14 - Scenario (f)

Default Path:

AHFG-10 either does not have a cache entry or does not have a shortcut for the destination internetworking address of H1. If no entry exists for the destination internetworking address, an entry is created for the purpose of flow detection. AHFG-10 performs an increment on the cache entry's "count" field and flow detection processing is performed. If a flow is not detected, the internetworking protocol packet is then passed to the DFFG with which the AHFG registered (DFFG-1). DFFG-1 then places the appropriate MAC header, determined as described in section 5.5.1, on the internetworking protocol packet and sends it to Edge Device-1 via LANE. The LEC and bridging components of Edge Device-1 are used to send the frame to the correct legacy port for H1.

Shortcut Path:

If AHFG-10 detects a flow to the internetworking address of H1, it sends an NHRP Request to ICFG-1 via ICCtl. ICFG-1 then sends an Egress Cache Imposition to Edge Device-1 and receives an Egress Cache Imposition Acknowledgment from Edge Device-1. The Egress Cache Imposition Acknowledgment indicates whether ED-1 can accept the additional shortcut. ICFG-1 sends an NHRP Response to AHFG-10 via ICCtl. AHFG-10 may then establish a shortcut VC to ED-1 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to H1, AHFG-10 encapsulates the internetworking protocol packet with the appropriate encapsulation for the shortcut. The packet is then sent to ED-1 using the VC specified in the Inbound Cache entry.

B3.1.2 Dual-Stack Host Flows

This section covers dual-stack host flows for the intra IASG scenarios as illustrated in Figure 15.

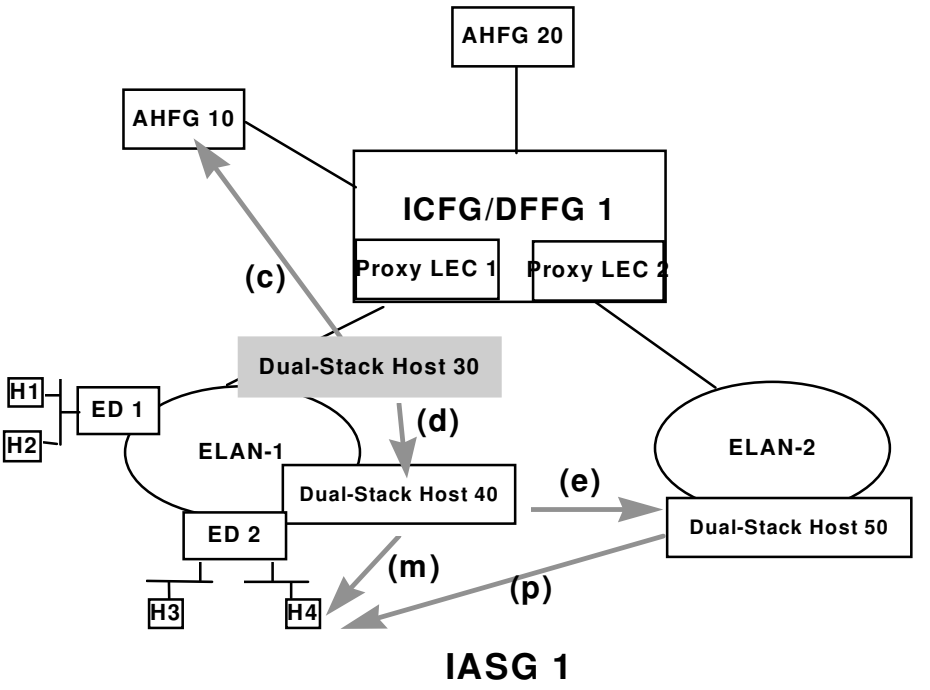


Figure 15 - Intra-IASG Dual-Stack Host Flows

B3.1.2.1 Scenario (c): Dual-Stack Host-30 to AHFG-10

For data originating from Dual-Stack Host-30 and destined to AHFG-10 within the same IASG, Figure 16 shows the default (LANE) and shortcut paths.

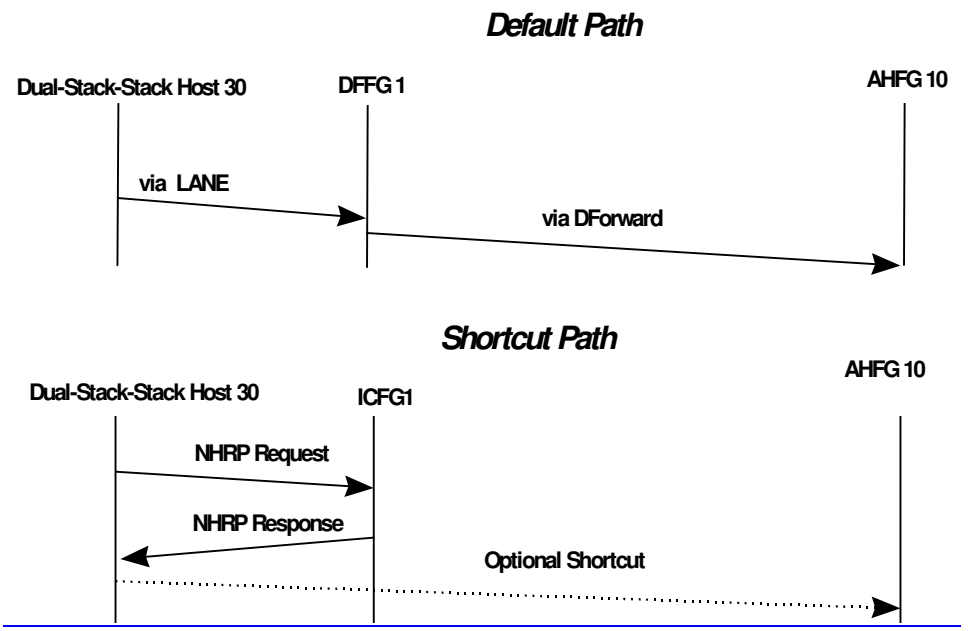


Figure 16 - Scenario (c)

Default Path:
The default LANE path shows Dual-Stack Host-30 using LANE to forward layer 2 frames to DFFG-1. DFFG-1 then removes the Layer 2 encapsulation and sends the internetworking protocol packet to AHFG-10 via DForward

Shortcut Path:
If Dual-Stack Host-30 detects a flow to the internetworking address of AHFG-10, it sends an NHRP Request to ICFG-1 via ICCtl. ICFG-1 then sends an NHRP Response to Dual-Stack Host-30 via ICCtl. Dual-Stack Host-30 may then establish a shortcut VC to AHFG-10 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to AHFG-10, Dual-Stack Host-30 encapsulates the internetworking protocol packet with the appropriate encapsulation for the shortcut. The packet is then sent to AHFG-10 using the VC specified in the Inbound Cache entry.

B3.1.2.2 Scenario (d): Dual-Stack Host-30 to Dual-Stack Host-40

For data originating from Dual-Stack Host-30 and destined to Dual Stack Host 40 within the same IASG, and ELAN, Figure 17 shows the default and shortcut paths.

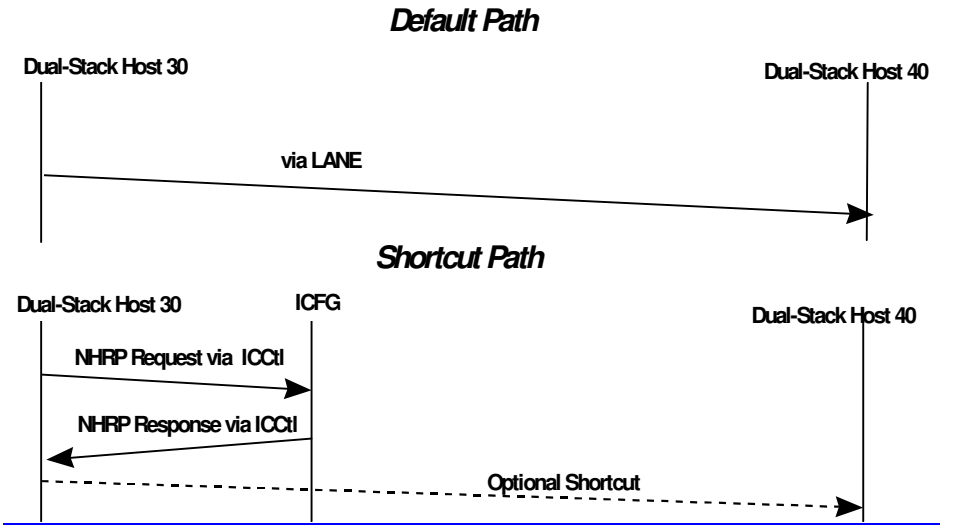


Figure 17 - Scenario (d)

Default Path
A LANE direct VC is used between these hosts for the default path..

Shortcut Path
This is the same as the shortcut path for scenario (c).

B3.1.2.3 Scenario (e): Dual-Stack Host-40 to Dual-Stack Host-50 (different ELANs)

For data originating from Dual-Stack Host-40 and destined to Dual-Stack Host-50 within the same IASG but a different ELAN, Figure TBD shows the default (LANE) data path.

Figure TBD

Default Path
LANE direct VCs are used between a bridge (currently outside the scope of MPOA) and the hosts.

Shortcut Path
There is no MPOA shortcut path for this scenario.

B3.1.2.4 Scenario (m): Dual-Stack Host-40 to Legacy Host H4

For data originating from Dual-Stack Host-40 and destined to legacy Host H4 within the same IASG, Figure 18 shows the data path.

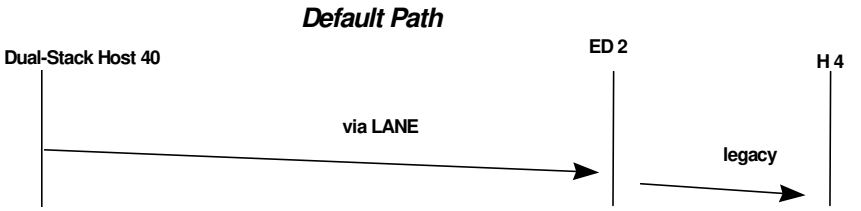


Figure 18 - Scenario (m)

Default Path:
This flow strictly uses LANE for all data transfer.

Shortcut Path:
There is no MPOA shortcut in this scenario.

B3.1.2.5 Scenario (p): Dual-Stack Host-50 to Legacy Host H4

For data originating from Dual-Stack Host-50 and destined to legacy Host H4 within the same IASG but a different ELAN, Figure 19 shows the data path.

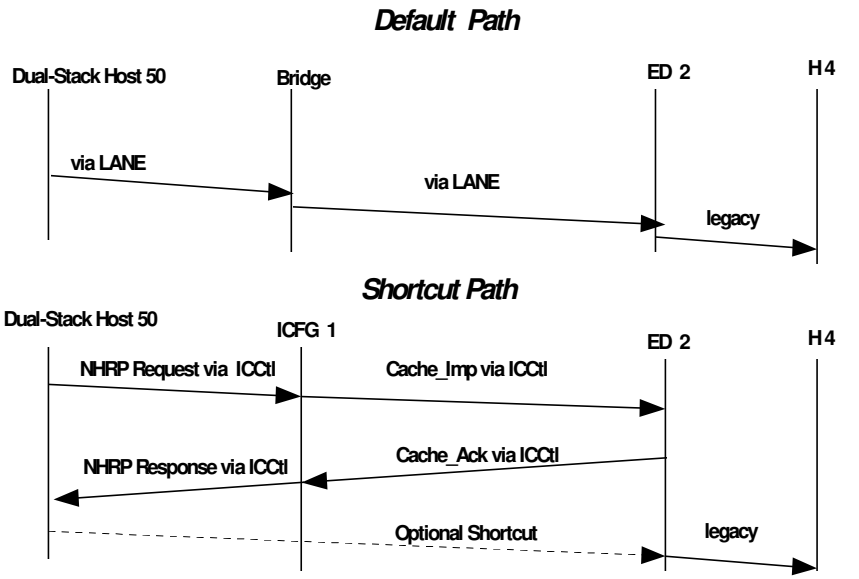


Figure 19 - Scenario (p)

Default Path:
LANE direct VCs are used between a bridge (currently outside the scope of MPOA) and the host as well as between a bridge and ED-2.

Shortcut Path:
The optional flow for this scenario is similar to scenario (f).

B3.1.3 Edge Device Flows

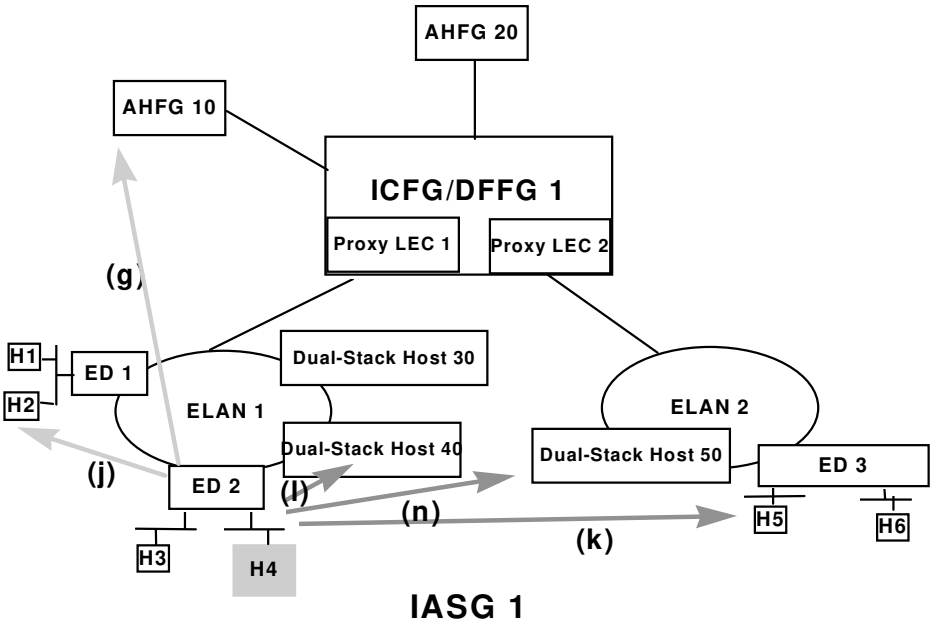


Figure 20 - Intra-IASG Edge Device Flows

B3.1.3.1 Scenario (g): Legacy Host H4 to AHFG-10

For data originating from Legacy Host H4 and destined to AHFG-10 within the same IASG, Figure 21 shows the default and shortcut data paths.

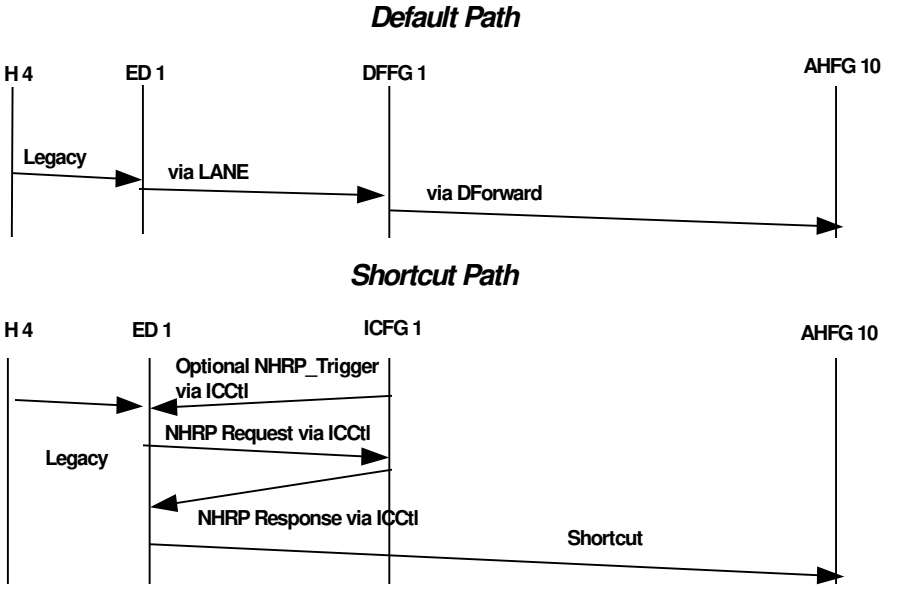


Figure 21 - Scenario (g)

Default Path:

The ED uses the destination MAC and the destination internetworking addresses as the key into the Inbound Cache. For this case, the Inbound Cache either does not have an entry for the key or has an

entry for the key but no shortcut exists. If no entry exists for the key, an entry is created in the Inbound Cache for the purpose of flow detection.

An increment is performed on the Inbound Cache entry's "count" field and flow detection processing is performed. A Layer 2 frame is then passed to the LEC of ED-1 which sends it to DFFG-1 via LANE. DFFG-1 then removes the Layer 2 encapsulation and sends the internetworking protocol packet to AHFG-10 via DForward.

Shortcut Path
If ED-1detects a flow to the internetworking address of AHFG-10, it sends an NHRP Request to ICFG-1 via ICCtl. ICFG-1 responds with an NHRP Response to ED-1via ICCtl. ED-1 may then establish a shortcut VC to AHFG-10 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to AHFG-10, ED-1 encapsulates the internetworking protocol packet with the appropriate encapsulation (see section 4.3.4) for the shortcut. The packet is then sent to AHFG-10 using the VC specified in the Inbound Cache entry.

Figure 21 also shows the optional NHRP Trigger which may be sent from ICFG-1 to ED-1, if ICFG-1 chooses to detect the flow.

B3.1.3.2 Scenario (l): Legacy Host H4 to Dual-Stack Host-40

For data originating from Legacy Host H4 and destined to Dual-Stack Host-40 within the same IASG, Figure 22 shows the data path.

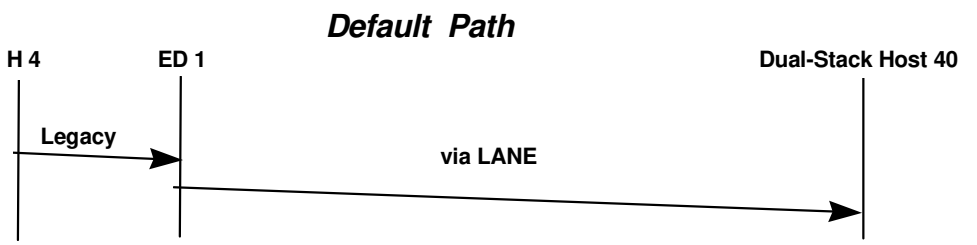


Figure 22 - Scenario (l)

Default Path:
This flow strictly uses LANE for all data transfer.

Shortcut Path:
There is no MPOA shortcut in this scenario.

B3.1.3.3 Scenario (n): Legacy Host H4 to Dual-Stack Host-50 (different. ELANs)

For data originating from Legacy Host H4 and destined to Dual-Stack Host-50 within the same IASG but a different ELAN, Figure 23 shows the data path.

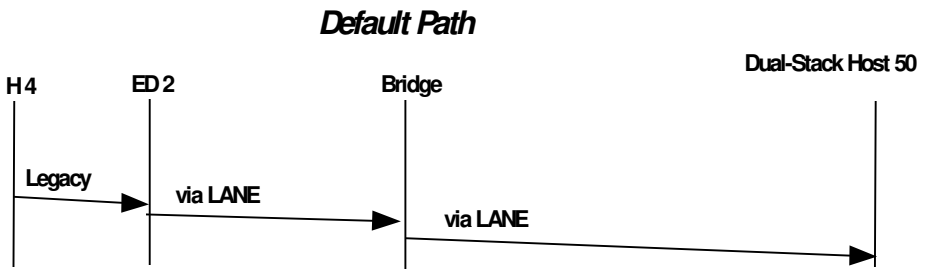


Figure 23 - Scenario (n)

Default Path:
LANE direct VCs are used between a bridge (currently outside the scope of MPOA) and the host as well as between a bridge and ED-1.

Shortcut Path:
There is no MPOA shortcut in this scenario.

B3.1.3.4 Scenario (j): Legacy Host H4 to Legacy Host H2

For data originating from Legacy Host H4 and destined to Legacy Host H2 within the same IASG, Figure 24 shows the data path.

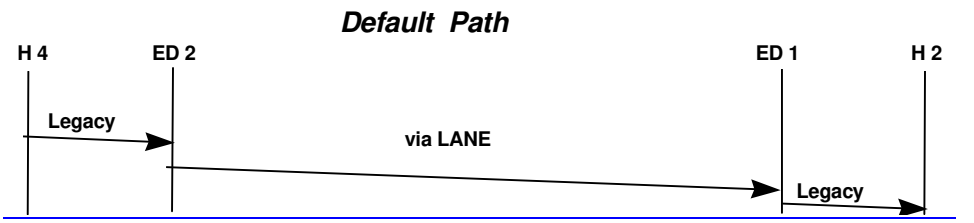


Figure 24 - Scenario (j)

Default Path:
Note that an MPOA shortcut path is not used. Rather, a LANE Direct VC is used. The ED component of ED-2 does not recognize the destination MAC as belonging to the RFFG or the DFFG. Therefore, the ED simply passes the MAC frame to the ED-2 LEC component for delivery to the ED-1. ED-2 LEC component receives the frame from the ATM network and passes it to the legacy bridging component of the edge device for delivery to the correct legacy port.

Shortcut Path:
There is no MPOA shortcut in this scenario.

B3.1.3.5 Scenario (k): Legacy Host H4 to Legacy Host H5 (different. ELANs)

For data originating from Legacy Host H4 and destined to Legacy Host H5 within the same IASG but a different ELAN, Figure 25 shows the data path.

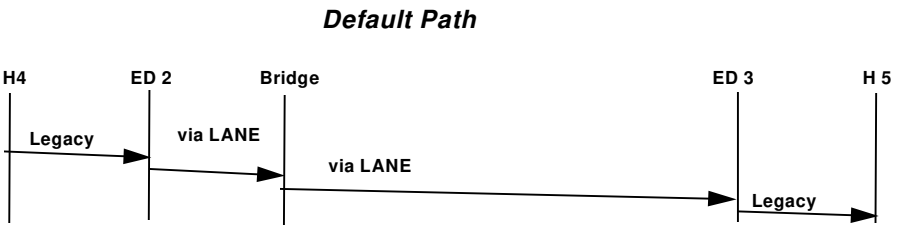


Figure 25 - Scenario (k)

Default Path:
LANE direct VCs are used between the EDs.

Shortcut Path:
There is no MPOA shortcut in this scenario.

B3.2 Inter-IASG

B3.2.1 AHFG Flows

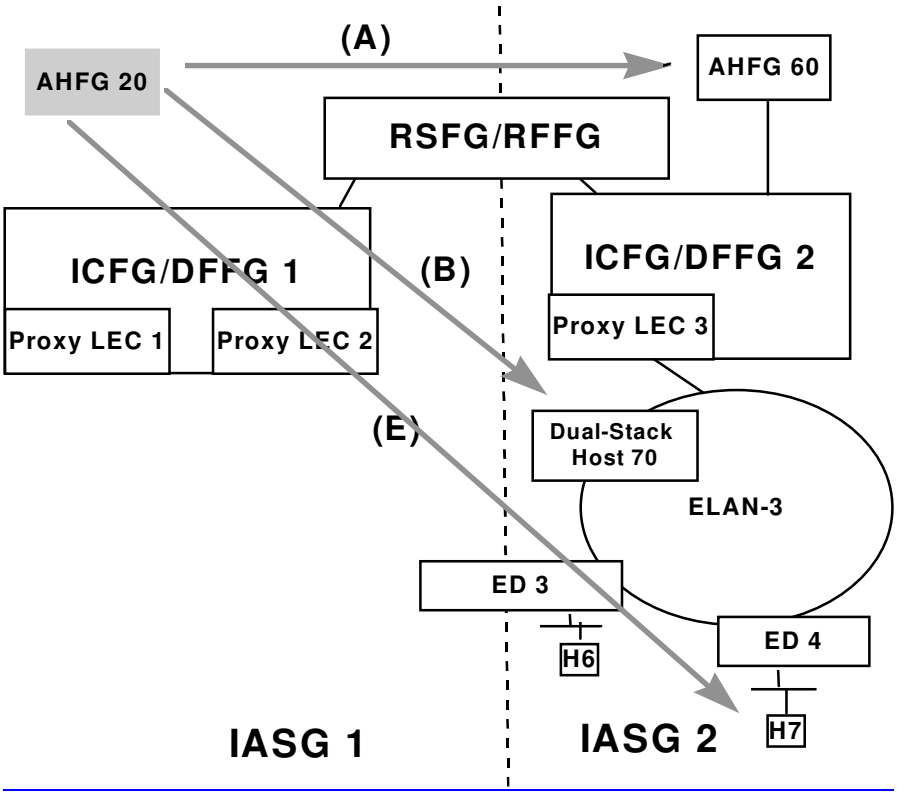


Figure 26 - Inter-IASG AHFG Flows

B3.2.1.1 Scenario (A): AHFG-20 to AHFG-60

For data originating from AHFG-20 and destined to AHFG-60 within a different IASG, Figure 27 shows the two possible default data paths.

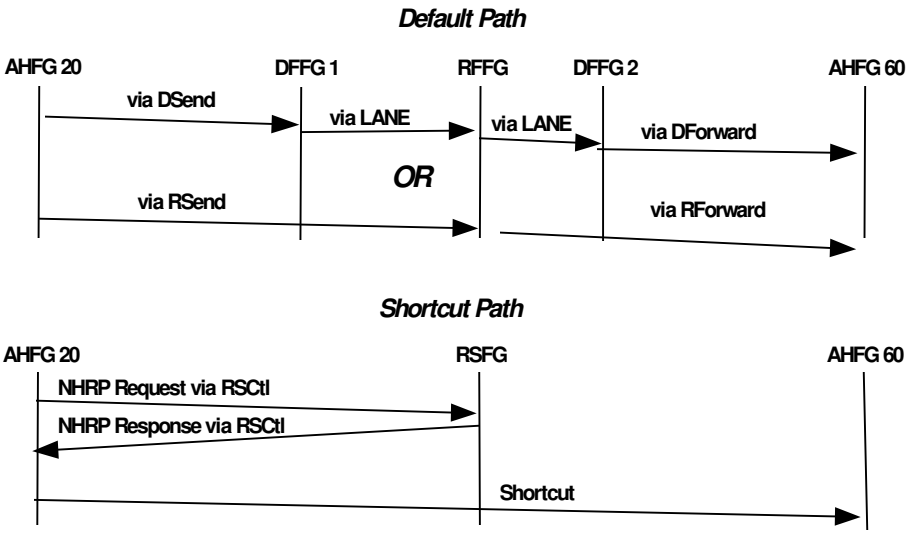


Figure 27 - Scenario (A)

Default Path:

In both cases, AHFG-20 updates the cache entry "count" field and performs flow detection processing. If AHFG-20 does not have the optional ability to determine if the source and destination are in different IASGs, AHFG-20 uses the DSend path to send the internetworking protocol packet to the DFFG with which it registered (DFFG-1). DFFG-1 encapsulates the internetworking protocol packet in layer 2 and sends the frame to the RFFG via LANE. The RFFG recognizes the destination internetworking address to be one associated with one of its "interfaces" and sends the layer 2 frame to DFFG 2 via LANE. DFFG-2 recognizes the destination internetworking address to be one associated with one of its registered AHFGs. DFFG-2 removes the layer 2 encapsulation and sends the internetworking protocol packet to AHFG-60 via the DForward path.

Also shown in Figure 27 is AHFG-20 performing the optional mask and match capability. AHFG-20 uses the RSend path to send the internetworking protocol packet directly to the RFFG. If the RFFG and DFFG-2 are co-resident, the RFFG just uses the RForward path to send the internetworking packet to AHFG-60.

Shortcut Path:

If AHFG-20 detects a flow to the internetworking address of AHFG-60, it sends an NHRP Request to ICFG-1 via ICCtl. ICFG-1 passes the request to the RSFG which sends an NHRP Response back to ICFG-1. ICFG-1 sends the NHRP Response to AHFG-20 via ICCtl. AHFG-20 may then establish a shortcut VC to AHFG-60 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to AHFG-60, AHFG-20 encapsulates the internetworking protocol packet with the appropriate encapsulation (see section 4.3.4) for the shortcut. The packet is then sent to AHFG-60 using the VC specified in the Inbound Cache entry.

If AHFG-20 performs the optional mask and match capability, it may send the NHRP Request directly to the RSFG via RSCtl.

B3.2.1.1 Scenario (B): AHFG-20 to Dual-Stack Host-70

For the default path, the scenario is similar to scenario (A) if dual host 70 registers with the DFFG 2. Otherwise, the path for DFFG 2 to Dual Host 70 goes through ELANs as depicted in scenario (E). The shortcut path is similar to that in scenario (B).

B3.2.1.3 Scenario (E): AHFG-20 to H7

For data originating from AHFG-20 and destined to Legacy Host H7 within a different IASG, Figure 28 shows the two possible default data paths.

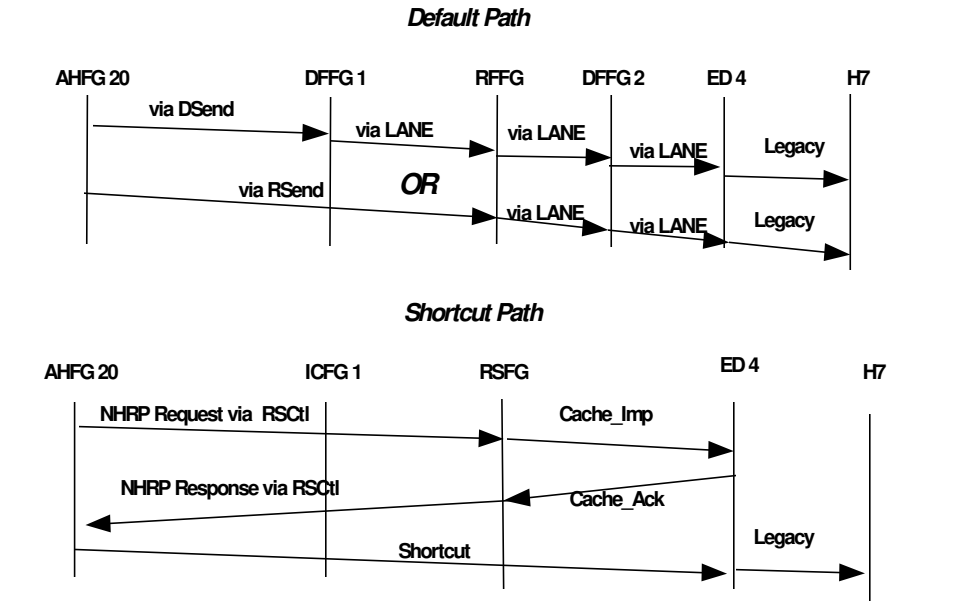


Figure 28 - Scenario (E)

Default Path:

In both cases, AHFG-20 updates the cache entry "count" field and performs flow detection processing. If AHFG-20 does not have the optional ability to determine if the source and destination are in different IASGs, AHFG-20 uses the DSend path to send the internetworking protocol packet to the DFFG with which it registered (DFFG-1). DFFG-1 encapsulates the interworking packet in Layer 2 and sends the frame to the RFFG via LANE.

The RFFG recognizes the destination internetworking address to be one associated with one of its "interfaces" and sends the layer 2 frame to DFFG 2 via LANE. DFFG-2, which determined the correct MAC header as described in section 6.3.3, updates the MAC header on the frame and sends it to ED-4 via LANE. ED-4 LEC and bridge components then send the frame out the correct legacy port.

Also shown in Figure 28 is AHFG-20 performing the optional mask and match capability. AHFG-20 uses the RSend path to send the internetworking protocol packet directly to the RFFG.

Shortcut Path:

If AHFG-20 detects a flow to the internetworking address of Legacy Host H7, it sends an NHRP Request to ICFG-1 via ICCtl. ICFG-1 passes the request to the RSFG which sends an Egress Cache Imposition to ED-4. The RSFG receives an Egress Cache Imposition Acknowledgment from ED-4 indicating whether ED-4 can accept the additional shortcut. The RSFG sends an NHRP Response back to ICFG-1. ICFG-1 sends the NHRP Response to AHFG-20 via ICCtl. AHFG-20 may then establish a shortcut VC to ED-4 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to Legacy Host H7, AHFG-20 encapsulates the internetworking protocol packet with the appropriate encapsulation (see section 4.3.4) for the shortcut. The packet is then sent to ED-4 using the VC specified in the Inbound Cache entry.

If AHFG-20 performs the optional mask and match capability, it may send the NHRP Request directly to the RSFG via RSCtl.

B3.2.2 Dual-Stack Host Flows

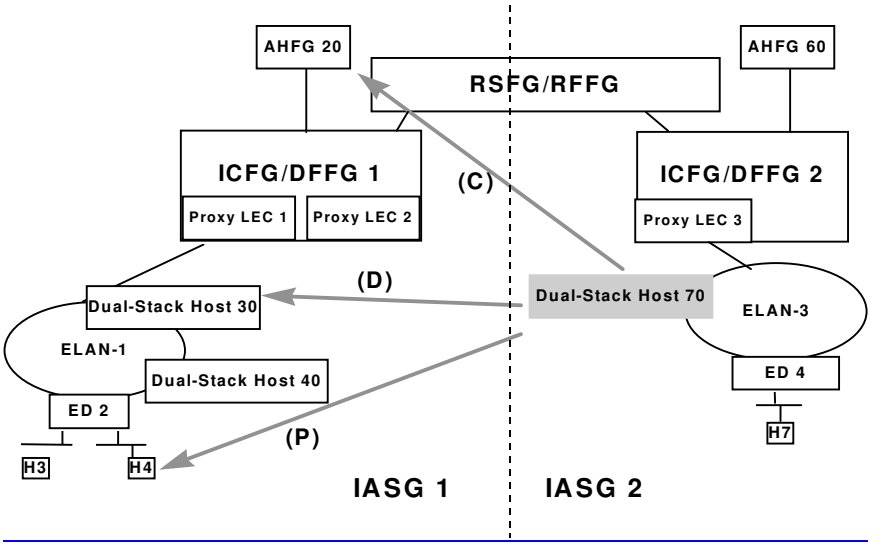


Figure 29 - Inter-IASG Dual Stack Host Flows

B3.2.2.1 Scenario (C): Dual-Stack Host-70 to AHFG-20

For data originating from Dual-Stack Host-70 and destined to AHFG-20 within a different IASG, Figure 30 shows the two possible default data paths.

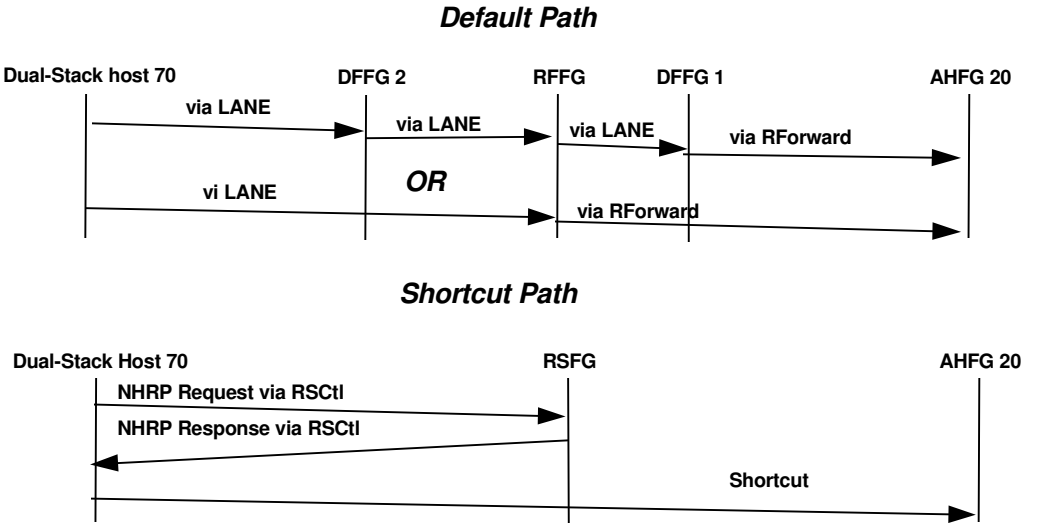


Figure 30 - Scenario (C)

Default Path:

In both cases, Dual-Stack Host-70 updates the cache entry "count" field and performs flow detection processing. If Dual-Stack Host-70 does not have the optional ability to determine if the source and destination are in different IASGs, Dual-Stack Host-70 uses LANE to send a layer 2 frame to the DFFG with which it registered (DFFG-2). DFFG-2 sends the layer 2 frame to the RFFG via LANE.

The RFFG recognizes the destination internetworking address to be one associated with one of its "interfaces" and sends the layer 2 frame to DFFG 1 via LANE. DFFG-1 recognizes the destination internetworking address to be one associated with one of its registered AHFGs. DFFG-2 removes the layer 2 encapsulation and sends the internetworking protocol packet to AHFG-20 via the DForward path.

Also shown in Figure 30 is Dual-Stack Host-70 performing the optional mask and match capability. Dual-Stack Host-70 uses the RSend path to send the internetworking protocol packet directly to the RFFG.

Shortcut Path:
If Dual-Stack Host-70 detects a flow to the internetworking address of AHFG 20, it sends an NHRP Request to ICFG-2 via ICCtl. ICFG-2 passes the request to the RSFG. The RSFG sends an NHRP Response back to ICFG-2. ICFG-2 sends the NHRP Response to Dual-Stack Host-70 via ICCtl. Dual-Stack Host-70 may then establish a shortcut VC to AHFG 20 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to AHFG 20, Dual-Stack Host-70 encapsulates the internetworking protocol packet with the appropriate encapsulation (see section 4.3.4) for the shortcut. The packet is then sent to ED-2 using the VC specified in the Inbound Cache entry.

If Dual-Stack Host-70 performs the optional mask and match capability, it may send the NHRP Request directly to the RSFG via RSCtl.

B3.2.2.2 Scenario (D): Dual-Stack Host-70 to Dual-Stack Host-30

For data originating from Dual-Stack Host-70 and destined to Dual-Stack Host-30 within a different IASG, Figure 31 shows the two possible default data paths.

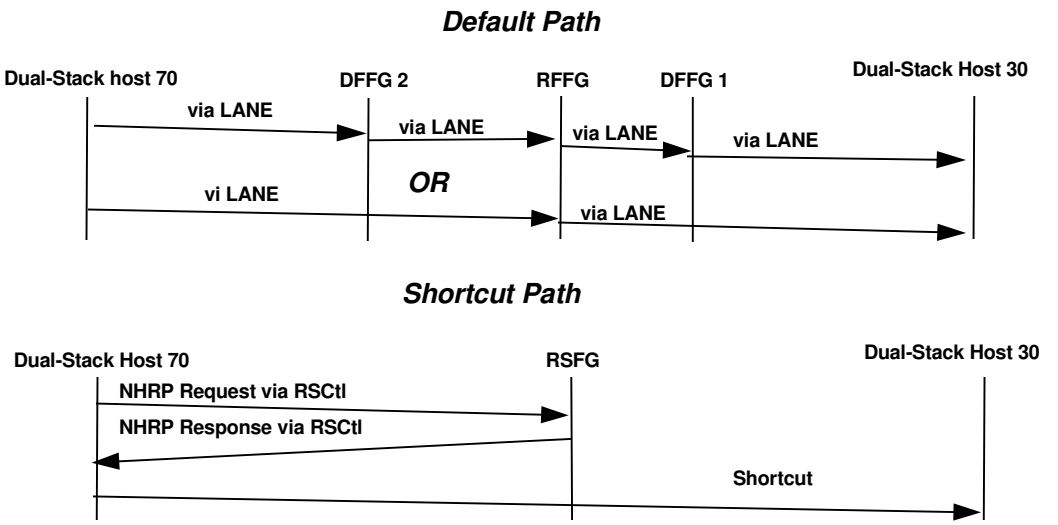


Figure 31 - Scenario (D)

Default Path:
In both cases, Dual-Stack Host-70 updates the cache entry "count" field and performs flow detection processing. If Dual-Stack Host-70 does not have the optional ability to determine if the source and destination are in different IASGs, Dual-Stack Host-70 uses LANE to send a layer 2 frame to the DFFG with which it registered (DFFG-2). DFFG-2 sends the layer 2 frame to the RFFG via LANE. The RFFG recognizes the destination internetworking address to be one associated with one of its "interfaces" and

sends the layer 2 frame to DFFG 1 via LANE. DFFG-1 sends a Layer 2 frame to Dual-Stack Host-30 via LANE.

Also shown in Figure 31 is Dual-Stack Host-70 performing the optional mask and match capability. Dual-Stack Host-70 uses LANE to send a layer 2 frame directly to the RFFG.

Shortcut Path:
If Dual-Stack Host-70 detects a flow to the internetworking address of Dual-Stack Host-30, it sends an NHRP Request to ICFG-2 via ICCtl. ICFG-2 passes the request to the RSFG which sends an NHRP Response back to ICFG-2. ICFG-2 sends the NHRP Response to Dual-Stack Host-70 via ICCtl. Dual-Stack Host-70 may then establish a shortcut VC to Dual-Stack Host-30 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to Dual-Stack Host-30, Dual-Stack Host-70 encapsulates the internetworking protocol packet with the appropriate encapsulation (see section 4.3.4) for the shortcut. The packet is then sent to Dual-Stack Host-30 using the VC specified in the Inbound Cache entry.

If Dual-Stack Host-70 performs the optional mask and match capability, it may send the NHRP Request directly to the RSFG via RSCtl.

B3.2.2.3 Scenario (P): Dual-Stack Host-70 to Legacy Host H4

For data originating from Dual-Stack Host-70 and destined to Legacy Host H4 within a different IASG, Figure 32 shows the two possible default data paths.

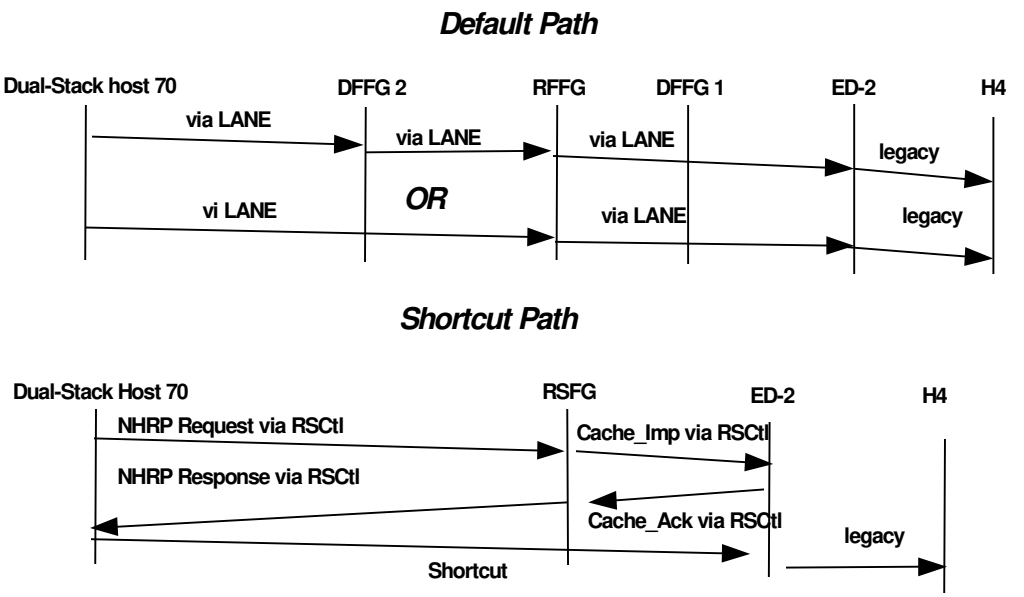


Figure 32 - Scenario (P)

Default Path:
In both cases, Dual-Stack Host-70 updates the cache entry "count" field and performs flow detection processing. If Dual-Stack Host-70 does not have the optional ability to determine if the source and destination are in different IASGs, Dual-Stack Host-70 uses LANE to send a layer 2 frame to the RFFG. The RFFG recognizes the destination internetworking address to be one associated with one of its "interfaces" and sends a layer 2 frame to ED-2 via LANE. ED-2 LEC then sends the frame out to the correct legacy port.

Also shown in Figure 32 is Dual-Stack Host-70 performing the optional mask and match capability. Dual-Stack Host-70 uses LANE to send a layer 2 frame directly to the RFFG.

Shortcut Path:

If Dual-Stack Host-70 detects a flow to the internetworking address of Legacy Host H4, it sends an NHRP Request to ICFG-2 via ICCtl. ICFG-2 passes the request to the RSFG which sends an Egress Cache Imposition to ED-2. The RSFG receives an Egress Cache Imposition Acknowledgment from ED-2 indicating whether ED-2 can accept the additional shortcut. The RSFG sends an NHRP Response back to ICFG-2. ICFG-2 sends the NHRP Response to Dual-Stack Host-70 via ICCtl. Dual-Stack Host-70 may then establish a shortcut VC to ED-2 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to Legacy Host H4, Dual-Stack Host-70 encapsulates the internetworking protocol packet with the appropriate encapsulation (see section 4.3.4) for the shortcut. The packet is then sent to ED-2 using the VC specified in the Inbound Cache entry.

If Dual-Stack Host-70 performs the optional mask and match capability, it may send the NHRP Request directly to the RSFG via RSCtl.

0.1.1 Edge Device Flows

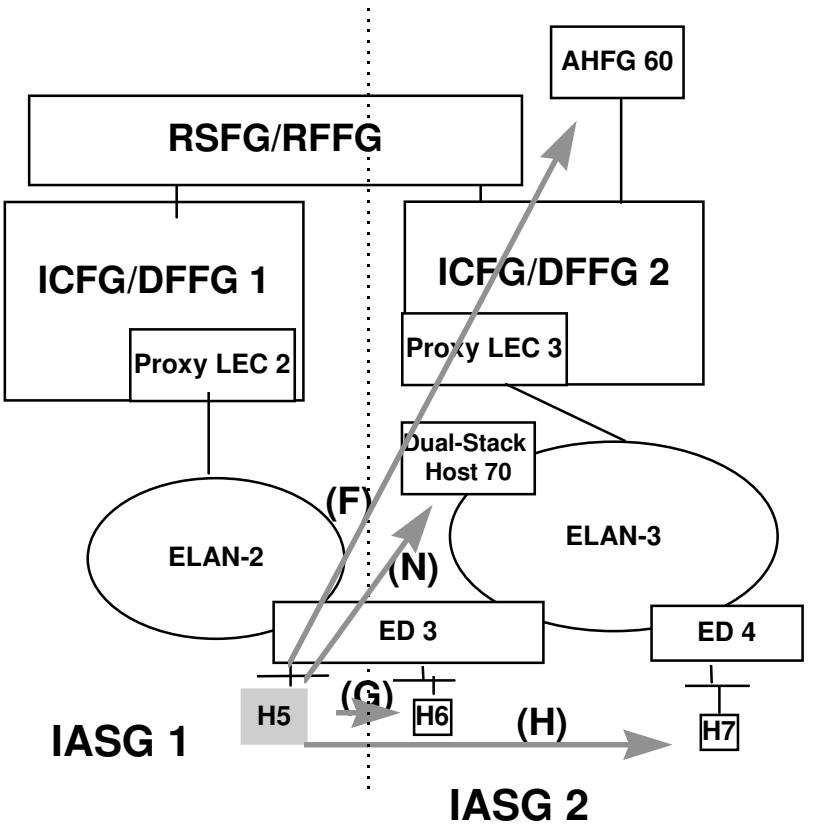


Figure 33 - Inter-IASG Edge Device Flows

B3.2.3.1 Scenario (F): Legacy Host H5 to AHFG 60

For data originating from Legacy Host H5 and destined to AHFG-60 some endpoint within a different IASG, Figure 34 shows the default and shortcut data paths.

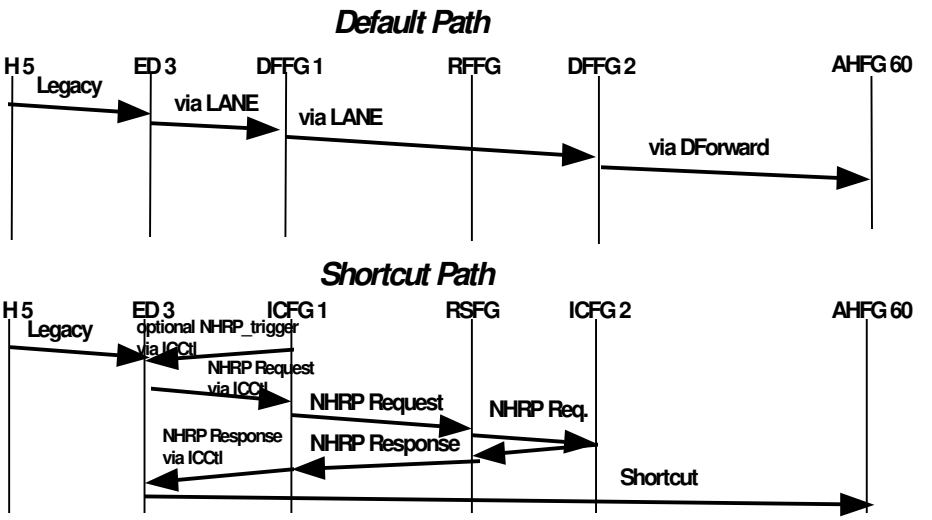


Figure 34 - Scenario (F)

Default Path:

ED-3 uses the destination MAC and the destination internetworking addresses as the key into the Inbound Cache. For this case, the Inbound Cache either does not have an entry for the key or has an entry for the key but no shortcut exists. If no entry exists for the key, an entry is created in the Inbound Cache for the purpose of flow detection. An increment is performed on the Inbound Cache entry's "count" field and flow detection processing is performed. ED 3 sends the layer 2 frame to the RFFG via LANE. The RFFG recognizes the destination internetworking address to be one associated with one of its "interfaces" and sends the layer 2 frame to DFFG 2 via LANE. DFFG-2 recognizes the destination internetworking address to be one associated with one of its registered AHFGs. DFFG-2 removes the layer 2 encapsulation and sends the internetworking protocol packet to AHFG-60 via the DForward path.

Shortcut Path:

If ED-3 detects a flow to the internetworking address of AHFG-60, it sends an NHRP Request to ICFG-1 via ICCtl. ICFG-1 sends the NHRP Request to the RSFG which associates the destination internetworking address with one of its local "interfaces". The RSFG sends the NHRP Request to ICFG-2 which responds with an NHRP Response. The RSFG passes the NHRP Response to ICFG-1 which also passes the NHRP Response to ED-3 via ICCtl. ED-3 may then establish a shortcut VC to AHFG-60 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to AHFG-60, ED-3 encapsulates the internetworking protocol packet with the appropriate encapsulation (see section 4.3.4) for the shortcut. The packet is then sent to AHFG-60 using the VC specified in the Inbound Cache entry.

Figure 34 also shows the optional NHRP Trigger which may be sent from the RSFG to ED-3 if the RSFG chooses to detect the flow.

B3.2.3.2 Scenario (G): Legacy H5 to Legacy H6, same Edge Device.

For data originating from Legacy Host H5 and destined to Legacy Host H6 within a different IASG but on the same Edge Device, Figure 35 shows the default and shortcut data paths.

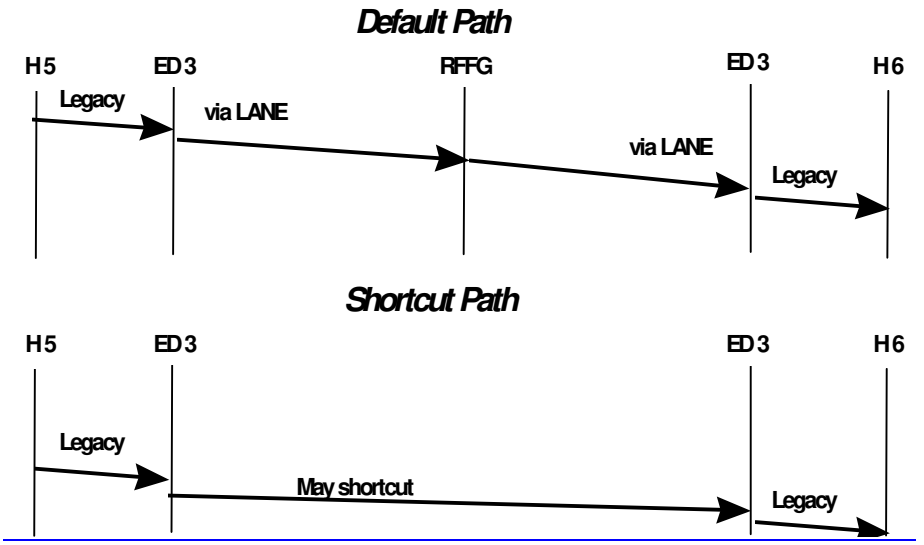


Figure 35 - Scenario (G)

Default Path:
ED-3 uses the destination MAC and the destination internetworking addresses as the key into the Inbound Cache. For this case, the Inbound Cache either does not have an entry for the key or has an entry for the key but no shortcut exists. If no entry exists for the key, an entry is created in the Inbound Cache for the purpose of flow detection. An increment is performed on the Inbound Cache entry's "count" field and flow detection processing is performed. A layer 2 frame is then passed to the LEC of ED-3 which sends it to the RFFG via LANE. The RFFG sends a layer 2 frame to DFFG-1 via LANE. DFFG-2 sends a layer 2 frame to ED-3. ED-3 LEC and bridge components then send the frame out the correct legacy port.

Shortcut Path:
If ED-3 detects a flow between Legacy Host H5 and Legacy Host H6, it is recommended that ED-3 internally route between the two attached hosts as described in section 6. However, ED-3 is permitted to establish a shortcut VC as shown in Figure 35.

B3.2.3.3 Scenario (H): Legacy Host H5 to Legacy Host H7

For data originating from Legacy Host H5 and destined to Legacy Host H7 within a different IASG, Figure 36 shows the default and shortcut data paths.

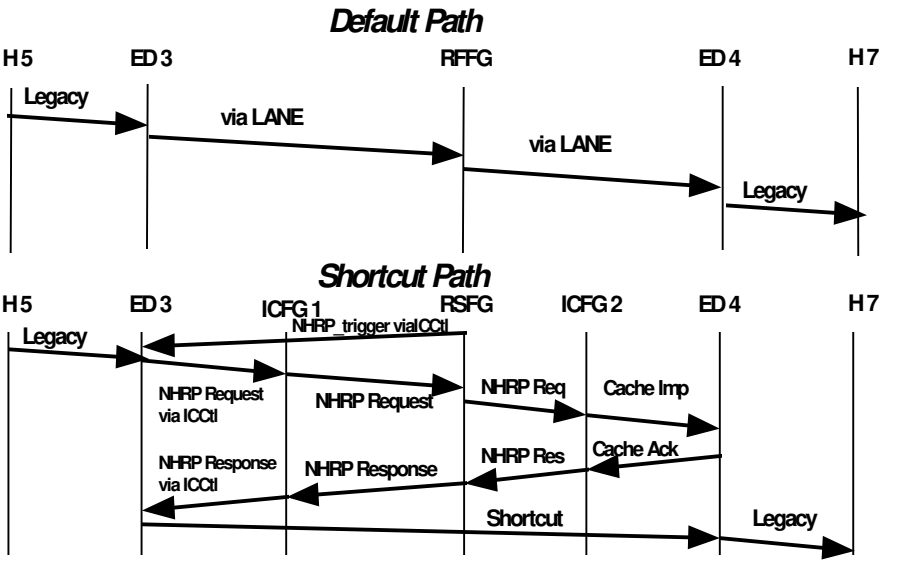


Figure 36 - Scenario (H)

Default Path:

ED-3 uses the destination MAC and the destination internetworking addresses as the key into the Inbound Cache. For this case, the Inbound Cache either does not have an entry for the key or has an entry for the key but no shortcut exists. If no entry exists for the key, an entry is created in the Inbound Cache for the purpose of flow detection. An increment is performed on the Inbound Cache entry's "count" field and flow detection processing is performed. A layer 2 frame is then passed to the LEC of ED-3 which sends it to the RFFG via LANE. The RFFG then sends a layer 2 frame to DFFG-2 via LANE. DFFG-2 sends a layer 2 frame to ED-4 via LANE. ED-4 LEC and bridge components then send the frame out the correct legacy port.

Shortcut Path:

If ED-3 detects a flow to the internetworking address of Legacy Host H7, it sends an NHRP Request to ICFG-1 via ICCtl. ICFG-1 sends the NHRP Request to the RSFG which associates the destination internetworking address with one of its local "interfaces". The RSFG sends the NHRP Request to ICFG-2 which sends an Egress Cache Imposition to ED-4. ED-4 responds with an Egress Cache Imposition Acknowledgment indicating whether it can accept the additional shortcut. ICFG-2 sends an NHRP Response to the RSFG which passes it to ICFG-1. ICFG-1 also passes the NHRP Response on to ED-3 via ICCtl. Edge Device-3 may then establish a shortcut VC to ED-4 using the information in the NHRP Response as well as update its Inbound Cache.

For subsequent data destined to Legacy Host H7, ED-3 encapsulates the internetworking protocol packet with the appropriate encapsulation (see section 4.3.4) for the shortcut. The packet is then sent to ED-4 using the VC specified in the Inbound Cache entry.

Figure 26 also shows the optional NHRP Trigger which may be sent from the RSFG to ED-3 if the RSFG chooses to detect the flow.

B3.2.3.4 Scenario (N): Legacy Host H5 to Dual-Stack Host 70

For data originating from Legacy Host H5 and destined to Dual-Stack Host-70 within a different IASG, Figure 37 shows the default data path.

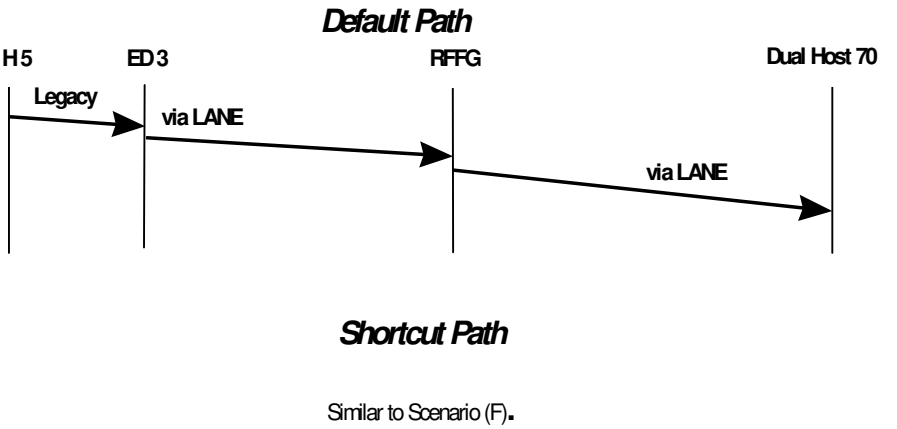


Figure 37 - Scenario (N)

Default Path:
ED-3 uses the destination MAC and the destination internetworking addresses as the key into the Inbound Cache. For this case, the Inbound Cache either does not have an entry for the key or has an entry for the key but no shortcut exists. If no entry exists for the key, an entry is created in the Inbound Cache for the purpose of flow detection. An increment is performed on the Inbound Cache entry's "count" field and flow detection processing is performed. A layer 2 frame is then passed to the LEC of ED-3 which sends it to the RFFG via LANE. The RFFG then sends a layer 2 frame to DFFG-2 via LANE. DFFG-2 sends a layer 2 frame to Dual-Stack Host-70 via LANE.

Shortcut Path:
This is similar to scenario (F).