SANDIA REPORT

SAND97–8273 • UC–405
Unlimited Release
Printed July 1997

A Survey of IP Over ATM Architectures

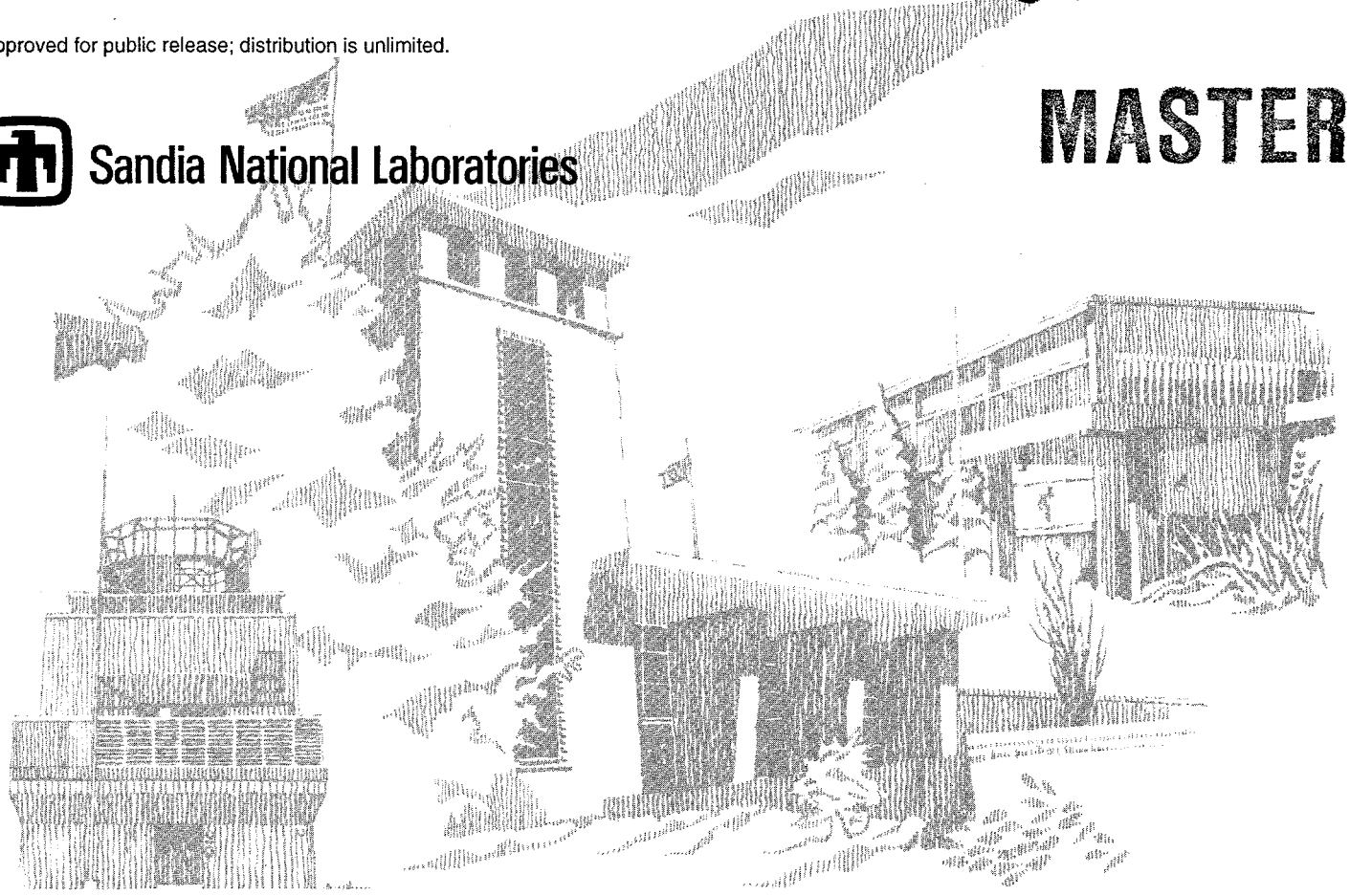
H. Y. Chen, R. Tsang, J. M. Brandt, J. A. Hutchins

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Approved for public release; distribution is unlimited.

Sandia National Laboratories



Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Prices available from (615) 576-8401, FTS 626-8401

Available to the public from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd
Springfield, VA 22161

NTIS price codes
Printed copy: A03
Microfiche copy: A01

SAND97-8273
Unlimited Release
Printed July 1997

UC-405

A Survey of IP over ATM Architectures

by Helen Chen, Rose Tsang, Jim Brandt, and Jim Hutchins
Infrastructure and Networking Research
Sandia National Laboratories, CA

Abstract

Over the past decade, the Internet has burgeoned into a worldwide information highway consisting of approximately 5 million hosts on over 45,000 interconnected networks. This unprecedented growth, together with the introduction of multimedia workstations, has spurred the development of innovative applications that require high speed, low latency, and real-time transport. Today's Internet can neither scale in its bandwidth nor guarantee the Quality of Services (QoS) necessary to meet these performance requirements. Many network researchers propose to use the Asynchronous Transfer Mode (ATM) [1] technology as the underlying infrastructure for the next generation of workgroup, campus, and enterprise IP networks. Since ATM is significantly different from today's legacy network technologies, efficient implementation of IP over ATM is especially challenging. This tutorial paper covers several existing proposals that integrate IP over ATM.

CONTENTS

1. Introduction 7

2. Background Information 7
   - Connectionless IP 7
   - The Integrated Services Extension to Internet 8
   - The ATM Technology 10

3. The IETF Approach 12
   - The Classical IP-over-ATM Model 12
   - The Next-Hop Resolution Protocol 14
   - The Multicast Address Resolution Service 15

4. The ATM Forum Approach 17
   - The LAN Emulation Protocol 17
   - The Multi Protocol over ATM Model 20

5. Proprietary Approaches 22
   - IP Switching 22
   - TAG Switching 24

6. Summary 27

1. Introduction

Over the past decade, the Internet has burgeoned into a worldwide information highway consisting of approximately 5 million hosts on over 45,000 interconnected networks. This unprecedented growth, together with the introduction of multimedia workstations, has spurred the development of innovative applications that require high speed, low latency, and real-time transport. Today's Internet can neither scale in its bandwidth nor guarantee the Quality of Services (QoS) necessary to meet these performance requirements. Many network researchers propose to use the Asynchronous Transfer Mode (ATM) [1] technology as the underlying infrastructure for the next generation of workgroup, campus, and enterprise IP networks. Aside from providing a transparent interface for best-effort traffic, these IP-over-ATM proposals also offer extensions to include future Integrated Internet Services [2]. This service architecture and its real-time protocols are being defined at the Internet Engineering Task Force (IETF). These new protocols will enhance the current Internet by allowing network resources to be allocated and guaranteed to real-time IP flows.

2. Background

Connectionless IP

Much of the Internet's success may be attributed to its simple transport paradigm which results in a remarkably robust and adaptive packet delivery system. As shown in Figure 1, the most fundamental service provided by the Internet architecture is a connectionless packet delivery system that uses the Internet Protocol (IP) [3]. This packet delivery system treats each packet independently from all others; routers in the path simply forward packets to the next-hop in the direction of the destination. This hop-by-hop packet forwarding paradigm allows IP to operate over a wide range of network technologies (e.g., Ethernet, token ring, FDDI, X.25, etc.). Furthermore, only soft states are required in routers; routing entries are refreshed periodically using routing protocols. In the event of link failures, routers simply update their routing table and send packets to an alternate route if it exists. This error recovery mechanism has proven to be extremely fault tolerant in the large Internet environment. While IP makes a best effort to deliver packets, there is no guarantee of reliable delivery. IP relies on an upper layer protocol such as the Transmission Control Protocol (TCP) [4] to provide end-to-end reliability .

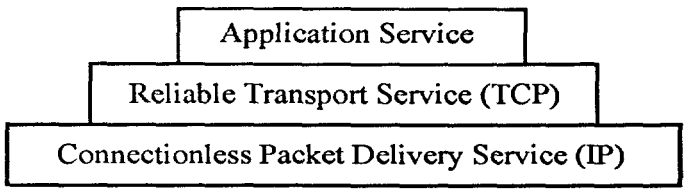


Figure 1. The Internet architecture.

The Integrated Services Extension to Internet

In order to meet the performance requirements of real-time applications, the IETF Integrated Internet Services working group is working on extending the current Internet architecture to include the support for real-time services. The core integrated service protocols consist of the next generation Internet Protocol (IPng) [5], the resource ReSerVation Protocol (RSVP) [6], the Real-time Transport Protocol (RTP) [7], and multicast routing protocols such as the Internet Group Management Protocol (IGMP) [8], the Distance Vector Multicast Routing Protocol (DVMRP) [9], the Protocol Independent Multicast (PIM) [10], and the Core Based Trees (CBT) protocol [11].

IP's multicast function is an integral part of new collaborative and real-time applications. When sending to multiple receivers, multicast minimizes bandwidth consumption as well as processing overhead. Moreover, if the destination address is unknown or changeable, multicast is a simpler and more robust alternative to unicast solutions that rely on directory servers, configuration files, or exhaustive search. Figure 2 presents the basic components of the IP multicast architecture. These components include the multicast host, multicast router, and the corresponding set of multicast routing protocols: 1) the host-to-router Internet Group Management Protocol (IGMP), and 2) the router-to-router Distance Vector Multicast Routing Protocol (DVMRP), Protocol Independent Multicast (PIM) protocol and Core Based Tree (CBT) protocol.

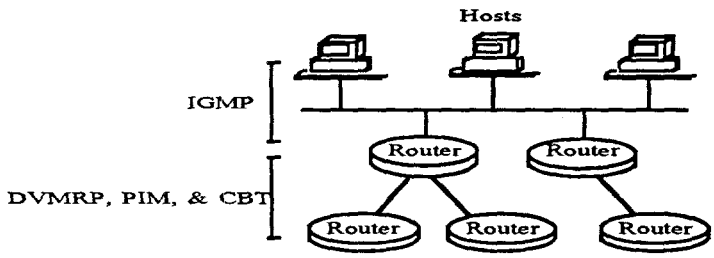
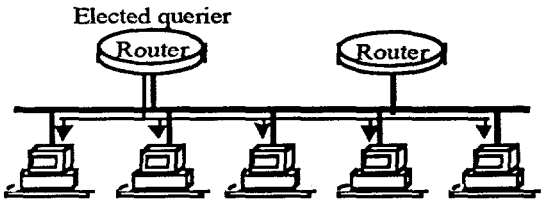


Figure 2. IP multicast components.

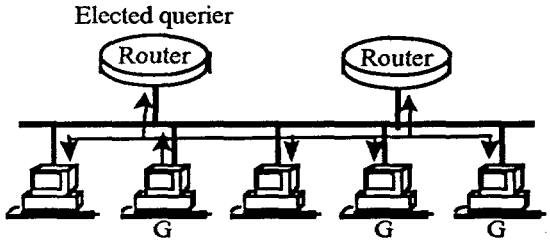
The IP multicast architecture specifies that: 1) each multicast group is identified by an IP group-address, 2) members of a group may be located anywhere in the Internet, and 3) members can join and leave at will. The IGMP protocol manages the IP hosts' group memberships using a reserved group-address (224.0.0.1). Currently, the IGMP protocol works over broadcast-capable LAN technologies such as Ethernet and FDDI. When a host decides to join a particular multicast group, it sends an IGMP-report message to a neighboring IGMP router. IGMP-report messages trigger multicast routers to update their group membership tables. In addition, a router will periodically send IGMP-query messages to solicit host reports in order that changes in membership activities on the link are reflected in a timely manner. As shown in Figure 3, random

delay timers in hosts ensure that only one copy of the IGMP-report message per group is sent to the querying router, thereby reducing bandwidth and processing overhead on the broadcast link.



- One querying router is elected per link
- The elected router sends an IGMP-query to solicit membership status using the allnode group address 224.0.0.1
- Hosts start random timers for each group to which they belong

(a)



- The host with the shortest delay timer sends an IGMP-report for group G to 224.0.0.1
- All other members hear and cancel their timer
- All routers hear the report and refresh their states
- Routers time out non-responding groups

(b)

Figure 3. The IGMP operations: a) query, and b) reply.

Most multicast routers forward multicast packets across routers using the simple Distance Vector Multicast Routing Protocol (DVMRP). Based on the routing information of a conventional distance-vector routing protocol such as the Routing Information Protocol (RIP) [12], each multicast router builds a shortest-path-tree (SPT) [13] (Figure 4a) using a subset of all routers such that all nodes within a multicast domain can be reached. The DVMRP protocol uses a data-driven mechanism to establish the eventual multicast network topology, where multicast packets are flooded through all routers along the SPT path (Figure 4b). Routers without active members send prune messages in order to be removed from the SPT (Figure 4c). Alternatively, if a router's subnet becomes active, it can join the SPT by sending a graft message (Figure 4d). To increase robustness, the flood/prune/graft procedures are repeated periodically so that all active members are reached through the SPT. As shown in Figure 4, the DVMRP protocol may waste

significant amounts of bandwidth as well as processing power, especially when group members are sparsely located across the entire network. Thus, this protocol will not scale well in the Internet environment. Core Based Trees (CBT) and Protocol Independent Multicast (PIM) have been proposed to address DVMRP's efficiency problems. These protocols reduce network resource overhead by sharing a multicast-tree among the sparsely populated group members. In addition, they use explicit join and leave messages instead of the flooding mechanism described above. Under the PIM scheme, routers with local members join toward a designated rendezvous-point (RP). Multicast sources forward packets to the RP which in turn multicast them to all active members throughout the multicast tree.

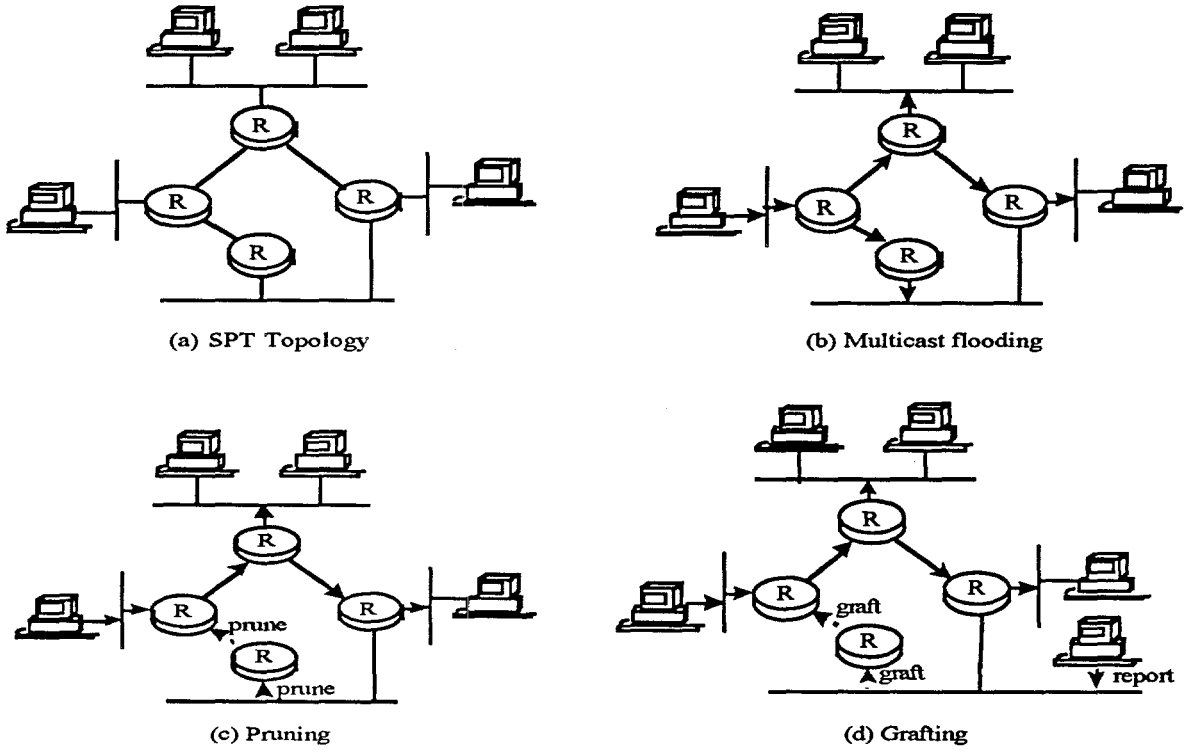


Figure 4. An example DVMRP topology.

The ATM Technology

Using cell-based switching technology, the Asynchronous Transfer Mode (ATM) can deliver important advantages such as flexible link speed, high aggregate bandwidth, and integrated services. Because an ATM switch schedules its transmission in units of 53-byte cells, as opposed to larger and/or variable size packets, ATM networks can offer better control in terms of end-to-end delay and delay-jitter. This unique ability enables ATM to guarantee Quality of Service (QoS) to a wide spectrum of traffic types such as audio, video, and data. Furthermore,

switching of fixed size packets simplifies ATM's VLSI implementation, thereby offering ATM switches the ability to potentially achieve throughput at the ever increasing link speeds. An ATM switch may achieve wire-speeds by implementing its data path completely in hardware. This requires connections to be pre-established. The connection setup process in ATM is depicted in Figure 5.

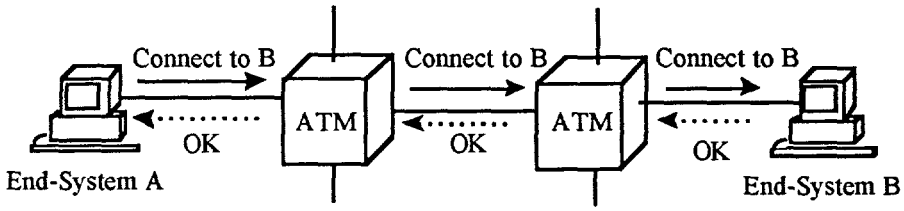


Figure 5. Connection setup through ATM signaling.

The steps of an ATM connection setup are: 1) node A sends a connect request using the ATM Q2931 signaling protocol [14] , 2) using the ATM Private Network to Network Interface (PNNI) [15] routing protocol, each ATM switch routes the connection request based on the Quality of Service requirements, 3) node B accepts the connection request, and the confirmation is propagated back along the connection path to node A. Any one of the ATM switches or the end system can reject the connection request if it does not have sufficient resources to satisfy the requested QoS. Consequently, ATM networks are connection-oriented; they require ATM switches to maintain hard states of pre-established connections in their forwarding tables.

There are two types of ATM connections (Figure 6): point-to-point and point-to-multipoint.

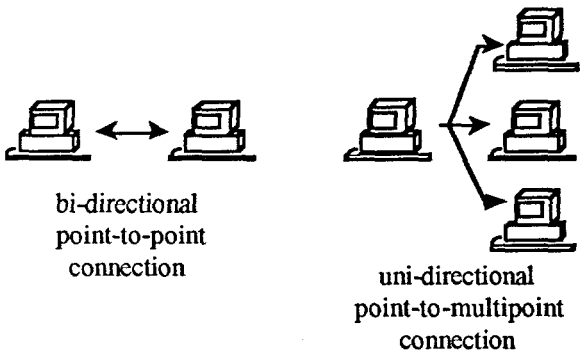


Figure 6. ATM connection types.

In an ATM point-to-multipoint connection, the 53-byte cell replication is done at the ATM switches where the branching occurs. Because cells from different packets, that are encapsulated in ATM Adaptation Layer 5 (AAL 5), are not allowed to interleave within a single connection, ATM point-to-multipoint connections are uni-directional only. Interleaving of cells from multiple packets within a single connection would preclude the reassembly of packets at the destination.

It is obvious that ATM is significantly different from existing networking technologies: 1) ATM is connection-oriented whereas IP uses a connection-less paradigm, and 2) ATM networks lack an analog to the multicast and broadcast capabilities that are inherent to shared medium technologies. In order to emulate shared medium multicast and broadcast, ATM networks need to support bi-directional multipoint-to-multipoint connections. However, such connections will cause interleaving of cells from multiple packets, thereby interfering with the packet reassembly processes at the destinations. Therefore, a simple and efficient integration of IP and ATM is especially challenging.

This tutorial paper provides a comparison of several existing proposals for supporting IP over ATM. In the following sections, we will describe the Classical IP-over-ATM [16] model as well as the ongoing Next-Hop Resolution Protocol (NHRP) [17] and the Multicast Address Resolution Protocol (MARS) [18] by the Internet Engineering Task Force (IETF), the Local Area Network Emulation (LANE) [19] and the ongoing Multiple Protocol Over ATM (MPOA) [20] work by the ATM Forum, the IP Switching [21] approach by IPSILON Networks (Palo Alto, California) , and the TAG Switching [22] solution by Cisco Systems (San Jose, California).

3. The IETF Approach

The Classical IP-over-ATM Model

The classical IP-over-ATM model, as defined by the IETF, uses ATM as the physical layer of the Internet protocol stack (Figure 7).

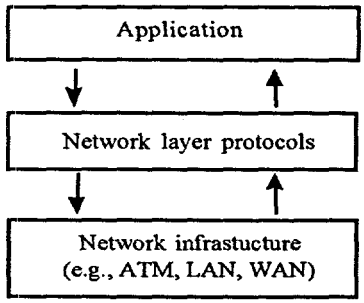


Figure 7. The Classical IP-over-ATM architecture.

This architecture allows the incorporation of the Integrated Internet Services which can provide packet level controls for accessing ATM level resources. While the Integrated Internet Services protocols offer applications a universal real-time network interface, the ATM infrastructure provides QoS guarantees. Figure 8 depicts the mapping of components between the Integrated Internet Services and the ATM infrastructure.

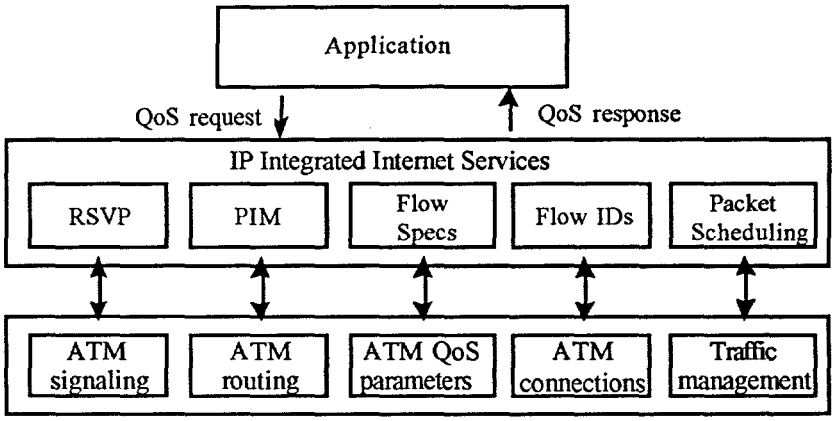


Figure 8. IP's Integrated Internet Services to ATM mapping.

As shown, the Classical IP-over-ATM (CLIP) approach completely decouples ATM's address space from the IP layer. Therefore, before IP nodes within an ATM-based IP subnet can communicate, they must first resolve each others' ATM addresses. This model defines the ATM Address Resolution Protocol (ATMARP) that will perform the IP to ATM address resolution using a server-based mechanism. Figure 9 depicts the basic CLIP procedure: 1) all nodes establish a connection with the ARP server-λ using the configured ATM address, 2) all nodes register their ATM as well as IP addresses with the designated ARP server, 3) node-α sends an ARP request to server-λ for node-β's ATM address, 4) node-α establishes an ATM connection to node-β using ATM's signaling protocol, and 5) node-α and -β proceed to communicate over this pre-established connection.

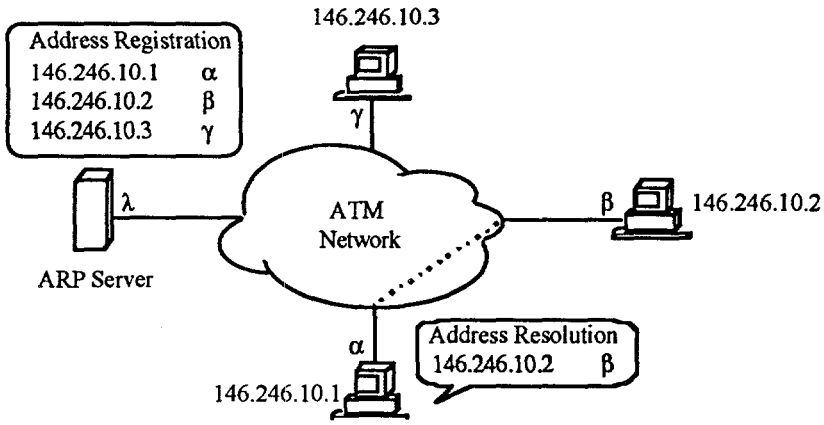


Figure 9. The ATMARP operations.

The Classical IP-over-ATM model follows the classical IP guideline which mandates that hosts on different IP subnets forward their packets via routers (Figure 10). Therefore, cut-through ATM connections are not allowed in this model, even when hosts are connected to the same ATM infrastructure. This restriction may cause performance degradation when the router becomes a bandwidth bottleneck and a single point of failure. Furthermore, CLIP lacks the capability to support IP multicast, which is undesirable because IP multicast is an integral part of most collaborative and real-time applications. The following sections will cover ongoing work at the IETF which address these issues.

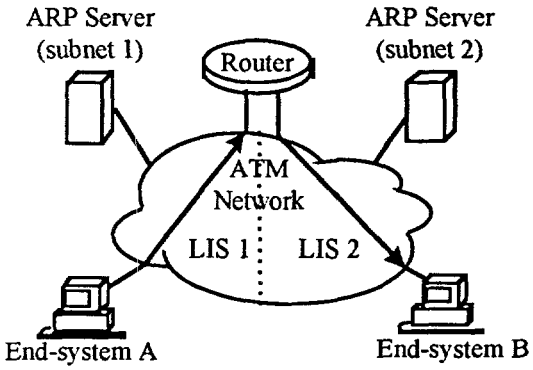


Figure 10. Routing in classical IP-over-ATM model.

The Next-Hop Resolution Protocol

The IETF Routing Over Large Cloud (ROLC) working group is working on the Next-Hop Resolution Protocol (NHRP) to overcome CLIP's inability to use cut-through paths. This model introduces a logical Non-Broadcast Multi-Access (NBMA) network which consists of a cluster of nodes that are connected to the same ATM infrastructure. The NHRP protocol allows cut-through paths to be established between nodes on the same NBMA network but belonging to

different logical IP subnets (LIS). A LIS is defined as the set of ATM nodes which share the same IP subnet address. Instead of ARP-servers, NHRP uses its own Next-Hop servers (NHS) to cache the next-hop IP-to-ATM address mappings. Typically, NHSs are the routers that serve ATM-attached nodes within the same LIS as well as legacy LANs which are reachable through these routers. Upon initialization, ATM attached systems register their IP and ATM addresses to their NHS. As shown in Figure 11, the NHRP supports communication between two NBMA nodes as follows: 1) the source node sends an NHRP-request, 2) the NHS returns the destination's ATM address if it is within the NHS's domain, 3) if not, the NHRP-request is forwarded to the next-hop NHS along the path, 4) steps 2 and 3 are repeated until a match of the IP address is found, 5) the corresponding ATM address is returned in an NHRP-reply message along the reverse path allowing all NHSs to cache the requested mapping, and 6) the source node sets up an ATM connection using the ATM signaling protocol and then proceeds to communicate directly with the destination node. While the NHRP-request is being processed, the NHRP protocol suggests that the source forwards packets along the default router path in order to improve perceived network response time.

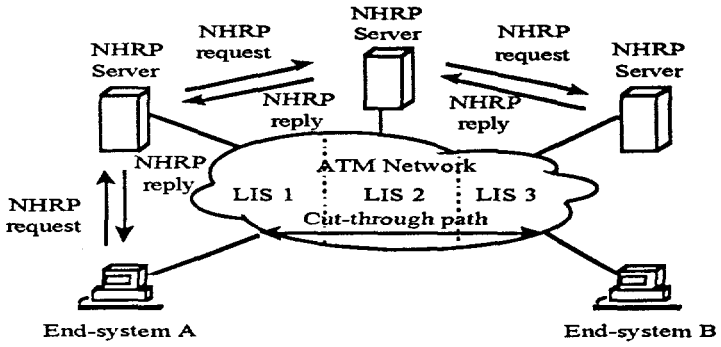


Figure 11. The NHRP operation

Because the cut-through route between two routers sends only data but not router adjacency information, the NHRP protocol violates a fundamental assumption of IP routing: routing updates are sent across all paths where data also flow. Therefore, persistent loops will result in multi-homed networks that have connections to both NHRP and legacy networks. Solutions to this problem are actively being pursued by the ROLC working group.

The Multicast Address Resolution Service

As mentioned earlier, CLIP's other limitation is its lack of IP multicast support. The ongoing Multicast Address Resolution Service (MARS) working group at the IETF proposes a model that adheres to the IP multicast architecture. In order to emulate the broadcast capability of a shared medium, this model adopts both a server-based and a mesh-based mechanism using ATM's point-to-multipoint connections. In the server-based approach (Figure 12a), multicast

nodes send their multicast packets to their multicast server via unidirectional point-to-point connections. These packets are subsequently forwarded to all group members via a unidirectional point-to-multipoint connection. The multicast server functions as a packet-sequencer, thereby preventing the interleaving of cells from different packets. In the mesh-based approach (Figure 12b), each sender has its own unidirectional point-to-multipoint connection. This set of point-to-multipoint connections in effect serves as a bi-directional multipoint-to-multipoint connection while also preventing cells from interleaving.

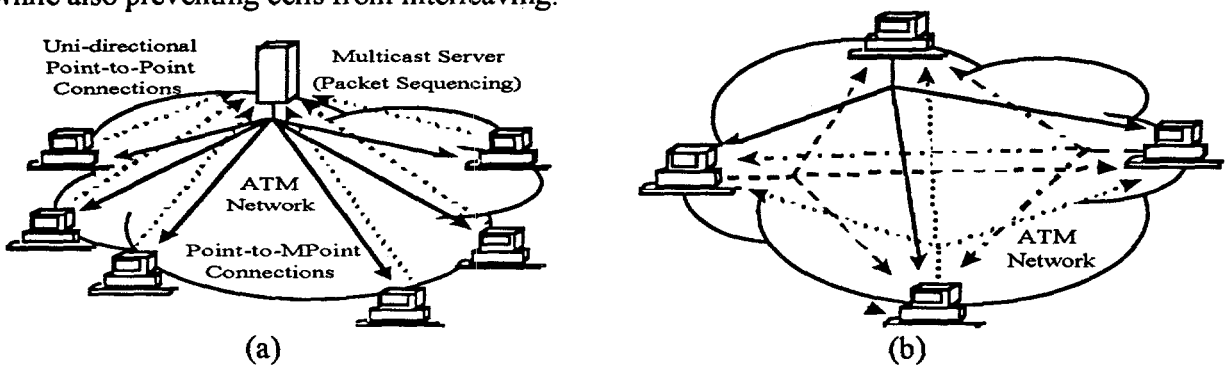


Figure 12. IP multicast in ATM networks: a) Server-based point-to-multipoint connection, b) mesh of point-to-multipoint connections.

A MARS server manages membership information for multicast groups within a region. Using a set of configuration control ATM connections (Figure 13), a MARS server gathers membership information as follows: 1) all nodes (i.e., hosts, routers, and multicast servers) are configured with the ATM address of their MARS server, 2) each node establishes a point-to-point connection to the MARS server which allows it to register for a multicast group, 3) in the server implementation, a multicast server uses this point-to-point connection to register the multicast group that it serves, 4) in return, the MARS server sends a list of this group's registered members, thereby allowing a point-to-multipoint connection to be established between the multicast server and its members, 5) in the mesh implementation, when a member wishes to transmit packets, it uses its point-to-point connection to retrieve group membership information from the MARS server, thereby allowing this node to establish a point-to-multipoint connection to other members, 6) the point-to-point connection is further used by all multicast nodes to transmit join or leave requests to the MARS server, 7) in order to convey join and leave requests, the MARS server establishes a point-to-multipoint Server Control Virtual Connection (VC) to all multicast servers and a point-to-multipoint Cluster Control VC to all multicast nodes, 8) appropriate multicast servers and/or hosts modify their point-to-multipoint connections to reflect the join and leave requests.

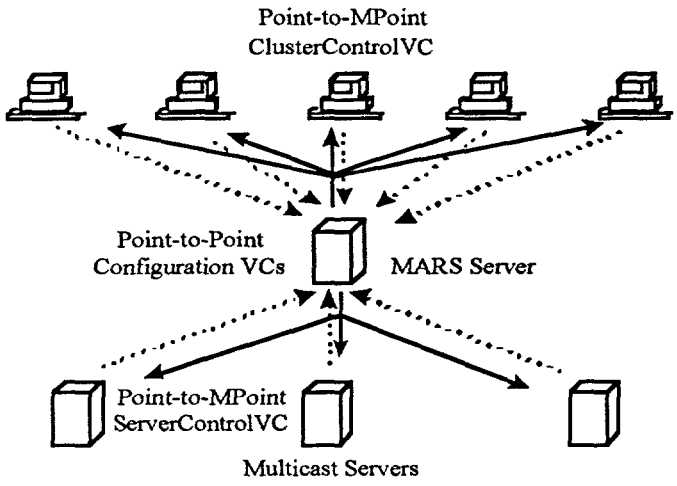


Figure 13. The MARS Configuration Control Connections.

The above mechanism, in effect, emulates the IGMP operations of the IP multicast. With the fundamental mechanisms in place, we now examine the actual multicast communication using MARS. In the server-based case, upon receiving a request to transmit from a node, the MARS server returns the ATM address of its multicast server. The requesting node then sets up a point-to-point connection to the multicast server and begins transmitting packets. The server in turn forwards these packets to all members using the newly established point-to-multipoint connection. In the case where the IP multicast address is served by a mesh of ATM point-to-multipoint connections, the MARS server will return the list of members to any member that wishes to transmit. Each transmitting member will then set up a point-to-multipoint connection for forwarding its own multicast packets.

While multicast packets crossing LIS boundaries can rely on routers that support IP multicast routing protocols such as Distance Vector Multicast Routing Protocol (DVMRP), Protocol Independent Multicast (PIM) or Core Based Trees (CBT), this mechanism, like the CLIP model, precludes the use of ATM cut-through paths. There are proposals at the IETF working group to extend PIM with NHRP support. The goal is to allow multicast point-to-multipoint connections to encompass ATM-attached members across LIS boundaries.

4. The ATM Forum Approach

The Local Area Network Emulation Protocol

Because of the large installed base of Ethernet and Token Ring LANs, the LAN Emulation (LANE) working group at the ATM Forum has proposed a protocol to provide interoperability

between ATM and these LAN technologies. The LANE protocol makes ATM appear like an Ethernet or Token Ring through a service interface which is identical to that of existing LANs. LANE encapsulates data in either an Ethernet or a Token Ring Medium Access Control (MAC) frame before sending it across the ATM infrastructure. After a LANE receiver recovers the MAC frame, it can either forward the frame to the higher layer protocols, if it is a host, or to a legacy LAN, if it is a layer-2 switch. Therefore, LANE hosts can communicate with other LANE hosts or legacy LAN hosts transparently (Figure 14).

Similar to IETF's approach, the LANE protocol also uses an overlay model which completely decouples the LANE protocol from its ATM infrastructure. Therefore, the LANE implementation requires an address resolution protocol that maps ATM addresses to MAC addresses. Like CLIP's ATMARP, LANE's address resolution is performed by a LAN Emulation Server (LES). The Broadcast and Unknown Server (BUS) supports broadcast and multicast, and provides a default path where the destination ATM address is unknown or connection setup is still in progress. A third server, the LANE Configuration Server (LECS), is used to assign Emulated LAN (ELAN) IDs and designated LESs for LANE clients in an administrative domain. Figure 15 depicts the major components of the LANE architecture.

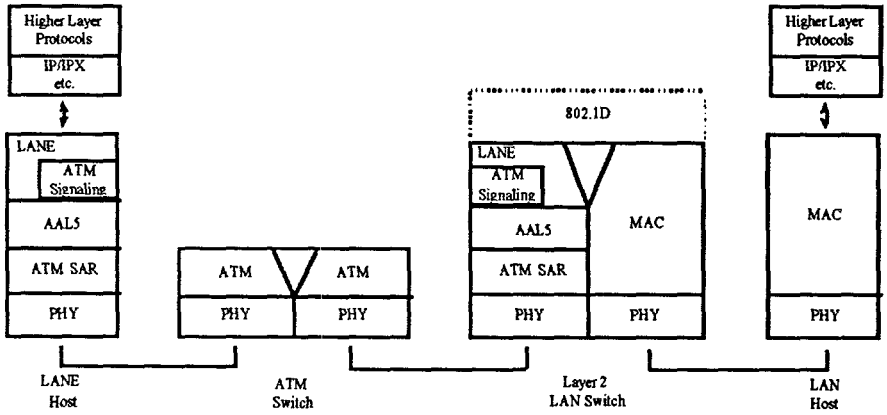


Figure 14. The LANE Architecture.

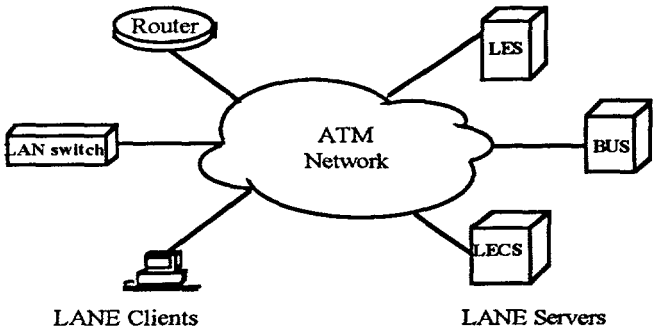


Figure 15. The LANE components.

The initialization process of a LANE client involves the establishment of five ATM connections (Figure 16). LANE clients acquire the ATM address of a LAN Emulation Configuration Server (LECS) via configuration, default connection, or auto-discovery. Using an ATM point-to-point Configuration-Direct connection, the LECS supplies the client with an ELAN ID and its LES ATM address. Through ATM signaling, the LANE client in turn sets up a point-to-point Control-Direct connection to the LAN Emulation Server in order to register its own ATM and MAC address pair. It is the LES's responsibility to add all registered clients to its point-to-multipoint Control-Distribute connection. The LES uses the Control-Distribute connection to solicit ATM addresses of unregistered clients such as hosts behind a LANE bridge. Finally a LANE client obtains its BUS's ATM address by sending a LANE ARP request using the MAC broadcast address of the emulated LAN. After obtaining the BUS ATM address, the LANE client can then set up a point-to-point Multicast-Send connection to the BUS and be on the BUS's point-to-multipoint Multicast-Forward connection.

LANE provides both unicast and broadcast communication using the control and data connections described in the previous paragraph. In the case of unicast communication, a LANE client must first resolve its partner's ATM address by issuing a LANE ARP request over the LES Control-Direct connection. With this ATM address, a point-to-point Data-Direct connection can then be established to carry its unicast traffic. The LANE protocol emulates broadcast and multicast by sending packets via the point-to-point Multicast-Send connection to the BUS. The BUS completes this function by forwarding these packets through its point-to-multipoint Multicast-Forward connection. These connections also serve as a default path for those packets that have unresolved ATM addresses, for instance, before a LANE client's LANE ARP request is completely processed. This alternative will improve the perceived network responsiveness by hiding the processing time of the LANE ARP request and the ATM signaling.

To summarize, the LANE model emulates LAN services across ATM networks using server mechanisms. As such, it provides interoperability with today's LAN hardware and software as

well as support for all their internetworking protocols. Using the LANE protocol, several ELANs can be implemented on the same ATM infrastructure, a technique which allows the construction of virtual LANs (VLAN). However, similar to the CLIP model, the LANE architecture also encounters performance limitations: 1) because the current version of the LANE protocol does not provide a mechanism to support distributed servers, these LANE servers can cause performance bottlenecks as well as single points of failure, and 2) communication between ELANs requires routers and, therefore, cannot take advantage of ATM's hardware speed. Furthermore, by emulating LAN interfaces, this solution suffers the same limitations as today's LAN technologies; it can not access ATM's QoS capabilities.

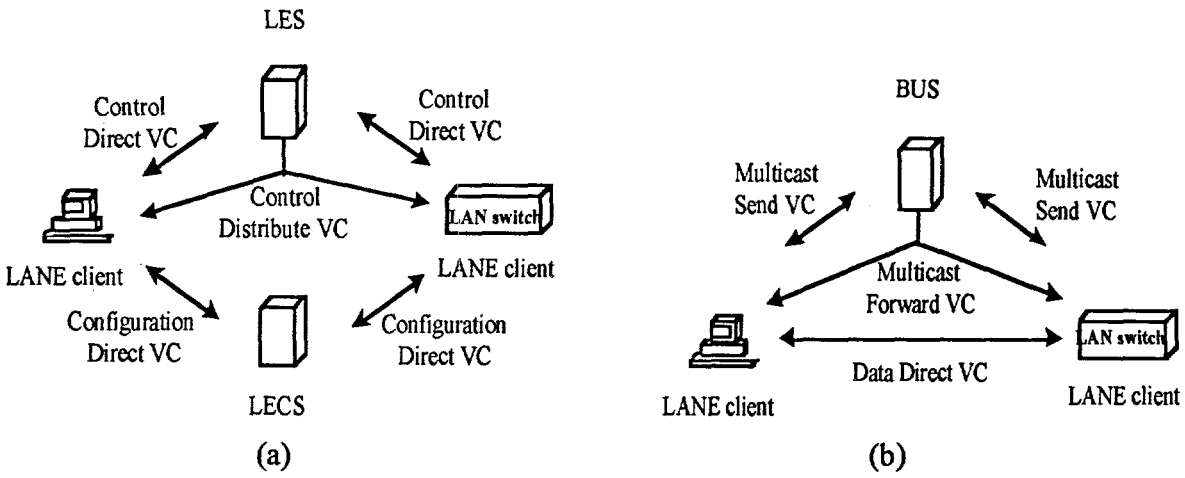


Figure 16. The LANE: a) control, and b)data connections.

Multi Protocol Over ATM

The Multi Protocol Over ATM (MPOA) working group at the ATM Forum is working on a solution that will preserve LANE efforts while allowing inter-subnet communications over direct ATM connections. The MPOA protocol achieves this goal by integrating LANE with NHRP, thereby providing the framework for the bridging and routing of diverse networking protocols and technologies over the ATM cut-through paths. The MPOA system extends the LANE protocol in order to provide MPOA devices with configuration services. Furthermore, the LANE ARP mechanism is employed by MPOA to allow auto-discovery among MPOA devices. Figure 17 depicts the relationships of the LANE and NHRP components within an MPOA system.

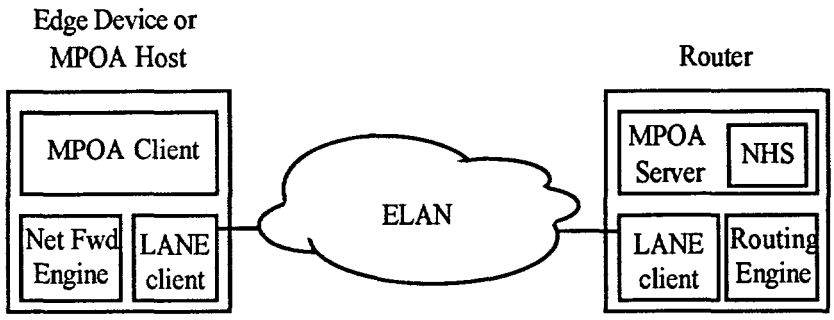


Figure 17. Components in MPOA system.

As shown in this figure, an MPOA system consists of an MPOA host or edge device, and an MPOA router. An MPOA device may function as either a network layer forwarding engine, in MPOA mode, or a MAC layer bridge, in LANE mode. As LANE clients, MPOA devices conduct intra-subnet communications as described in the previous section. When communication crosses sub-network boundaries, MPOA forwards packets, initially, along the default hop-by-hop router path (Figure 18a). When the traffic flow exceeds a pre-configured threshold on an MPOA client, the MPOA client will attempt to set up a short-cut path by sending a target-resolution request to its MPOA server. If the MPOA server is unable to resolve the target ATM address, the NHS component will transmit a NHRP request to its next-hop NHS. This process continues until the target ATM address can be resolved. The last-hop NHS then sends an MPOA imposition request to the destination MPOA client to verify that it has sufficient resources for this connection. If so, the last-hop NHS originates an NHRP reply which allows the target ATM address to be propagated back to the originating NHS. At this point, the first-hop MPOA server returns the resolved ATM address to the requesting MPOA client, thereby allowing a short-cut path to be established. The MPOA short-cut operations are illustrated in Figure 18b.

In an attempt to also provide short-cut paths to multicast traffic, the MPOA working group is working on extending the NHRP protocol to multicast routing protocols such as the Protocol Independent Multicast (PIM). The intent is to extend the ATM point-to-multipoint connection to multicast receivers that cross logical subnet boundaries.

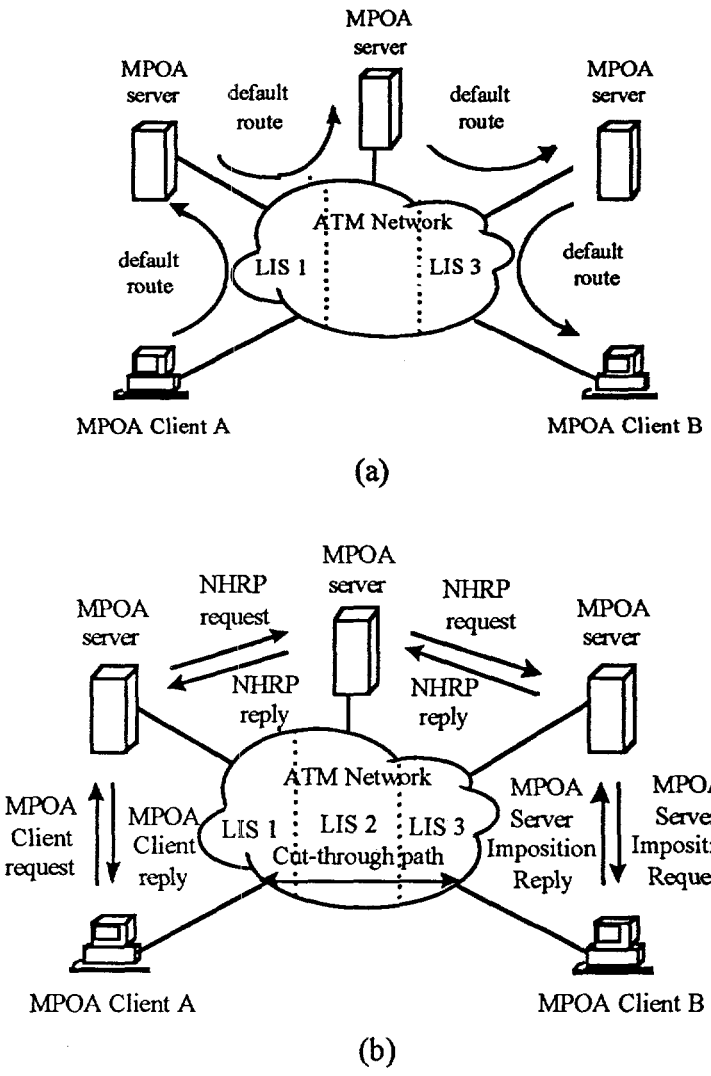


Figure 18. The MPOA operation: a) default hop-by-hop path, b) the short-cut path.

5. Proprietary Approaches

IP Switching

All of the overlay models mentioned require complicated protocols in order to emulate certain layer-2 or layer-3 services. In most cases, these protocols mandate a server-based implementation which invariably introduces performance bottlenecks and single points of failure. Furthermore, these models require the duplication of address spaces as well as routing algorithms which add administrative overhead and complexity in trouble-shooting. To IP, the ATM infrastructure appears as an opaque cloud and all ATM attached routers are adjacent to one

another regardless of the physical topology. In large Internet environments, this presents a scalability issue because a router's route processing and table requirements increase as N^2 where N is the number of edge routers. Presently, there are still unresolved technical issues such as possible occurrences of persistent loop in the NHRP algorithm.

IPSILON's IP Switching is one alternative to the previously discussed overlay models. This approach discards the connection oriented nature of ATM and integrates the fast ATM hardware directly with IP, thereby preserving IP's connectionless paradigm. The IP Switching system consists of an IP Switch and an IP Gateway/Host. Figure 19 depicts the software and hardware components of an IP switch.

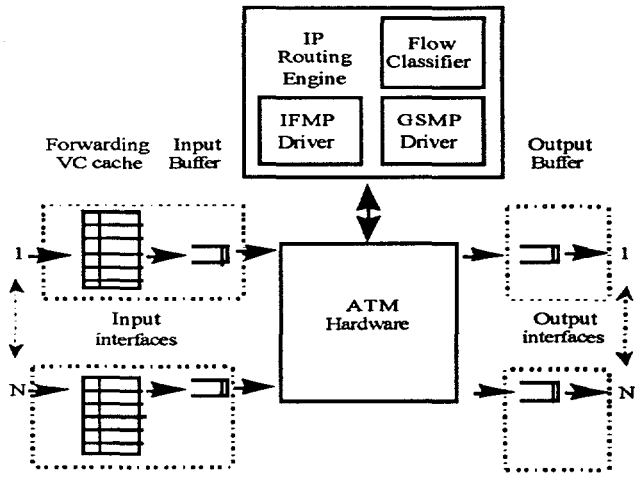


Figure. 19 The IP Switch components.

The IP Switch approach replaces the ATM software (e.g., ATM Q2931 signaling and ATM PNNI routing) in the switch control processor with a standard IP routing package. In addition, a driver based on IPSILON's General Switch Management Protocol (GSMP) manages cut-through ATM connections. By default, all IP traffic is forwarded hop-by-hop via the default ATM connection (Figure 20a). Meanwhile, a Flow Classifier, which resides both in the IP Switch and IP Gateway, monitors the IP traffic based on the local policy. If an IP flow qualifies for a short-cut path, the GSMP driver allocates a unique label for its short-cut ATM connection. After this short-cut path is in place, an IFMP (IPSILON Flow Management Protocol) redirect message is sent to its previous-hop IPSILON Switch or Gateway for switching from the default route to the short-cut path (Figure 20b). GSMP refreshes the switch connection table periodically in order to time out idle or dead connections. By maintaining soft states in the table, the IP Switching protocols are kept simple and robust; these protocols rely on time-outs to handle most of the error conditions.

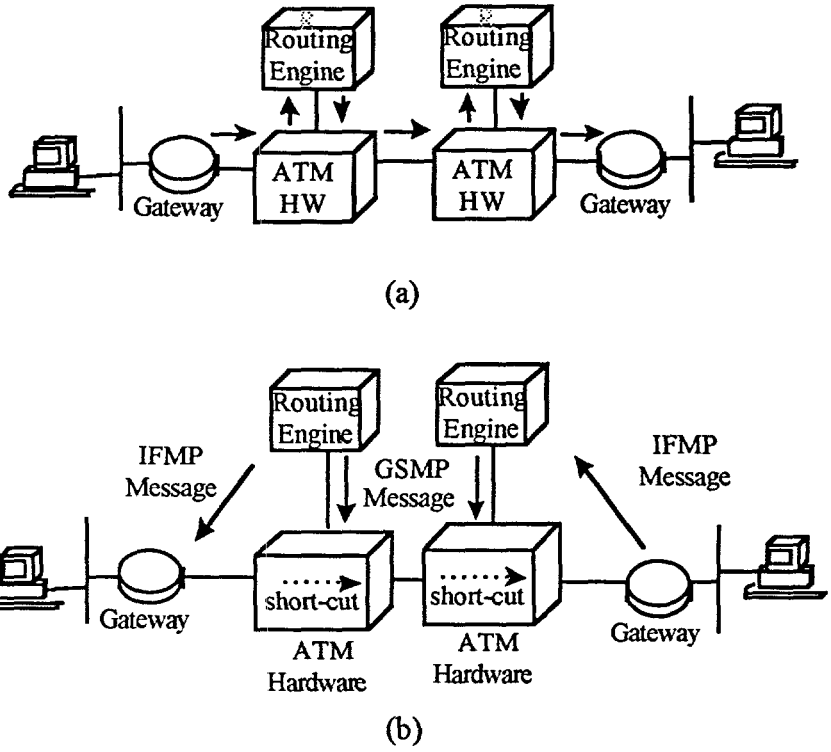


Figure 20. IP Switch connections: a) default and b) cut-through.

As shown, IP Switching proposes a point-to-point network model as opposed to the shared-medium model emulated by CLIP and LANE. Therefore, with IP Switching there is no need to emulate shared-medium broadcast using servers and complicated protocols. Since unicast and multicast routing protocols already work with point-to-point links, IP Switching can support IP multicast without modifications to the IGMP or the PIM protocol. IFMP's multicast-flow redirection works identically to its unicast flows. Although the current IP Switching implementation only supports a priority based QoS service, in the future, IPSILON plans to provide dynamic QoS guarantees through RSVP and GSMP operations. QoS guarantees will be supplied via scheduling and queue control within the ATM switch. IPSILON has submitted their IP Switching architecture and protocol descriptions as RFC documents to the IETF.

TAG Switching

Cisco has proposed a novel approach, called TAG Switching, which will improve the scaling properties of the Internet's routing systems. While this technology was originally designed for packet switching, Cisco will extend it to support ATM technology in order to provide a simpler

alternative to the overlay models described in the previous sections. Cisco's TAG Switching architecture integrates layer-3 routing with layer-2 switching. Since its layer-2 switching is based on small fixed-sized tags, this function can be implemented completely in hardware and thus benefit from significant performance gains. The TAG Switching architecture consists of two main components: control and forwarding. The forwarding component forwards packets using the tag-information carried in packets and the information kept in the tag-forwarding information base (TIB). The control component maintains tag forwarding information among interconnected TAG switches.

TAG Switching's forwarding paradigm is based on label swapping which uses the tag value in a packet as well as the information of a matching entry in the TIB. Each entry in the TIB consists of an incoming-tag, an outgoing-tag, the output interface, and the corresponding link level information (e.g., MAC address). Figure 21 gives an example of the tag forwarding operation: 1) TAG Switch-1 indexes into its TIB using the tag value of the incoming packet, 2) TAG Switch-1 swaps the packet's tag value from 3 to 9, 3) TAG Switch-1 forwards the packet to its output interface, fddi-0, 4) TAG Switch-2 indexes into its TIB, 5) TAG Switch-2 swaps the tag value from 9 to 2, and 6) TAG Switch-2 forwards packet to its output interface, eth-1.

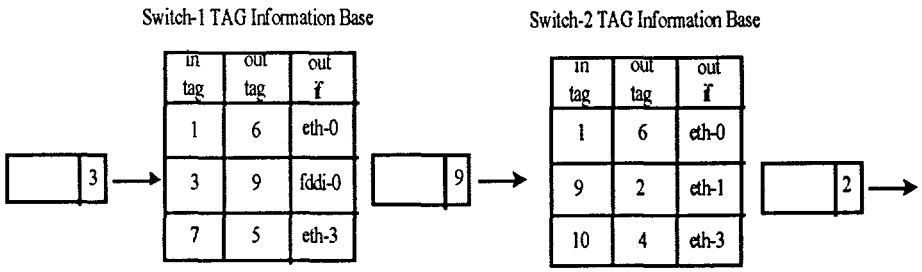


Figure 21. The label swapping paradigm used in TAG Switching.

Since TAG Switching forwarding decisions are based on network layer routing, it is essential that tags are associated with accurate network reachability information. The control component is responsible for creating and distributing tag-bindings among interconnected TAG Switches. Tag creation and distribution are driven by network topology. Typically, a TAG Switch creates and binds a tag to each routing entry in the routing table. These bindings, a tag value and a route, are then distributed to all neighbors that are directly connected. Upon receipt of a binding message, a neighboring TAG Switch will add the tag only if the associated route is downstream from it. This tag will be added to the TIB as an output tag, and the interface, where this binding was received, as the output interface. Since routes with a common prefix can be aggregated, a tag can be shared by multiple IP flows. Therefore, TAG Switching is as scaleable as current routing protocols. TAG Switching also implements stacks of tags where tags can be pushed or popped in accordance with layered routing information of the hierarchical IP routing model [24].

Since ATM forwarding is also based on label swapping, Cisco claims that TAG Switching can be readily applied. To support TAG Switching, the ATM switch must run the standard routing protocols as well as the TAG Switching control component. Under this scenario, tags will be carried in the Virtual Circuit Identification (VCI) field of the ATM header (Figure 22).

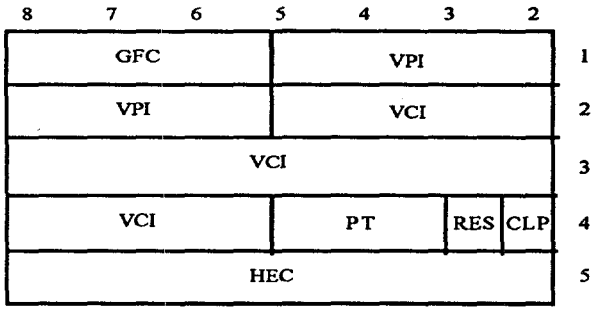


Figure 22. ATM header.

However, ATM imposes the following limitations to TAG Switching: 1) the ATM header lacks the necessary format to carry more than two layers, (VPI, VCI) (Figure 22), of tags which prevents it from fully supporting IP's hierarchical routing architecture, and 2) ATM TAG Switching cannot support route aggregation; a shared tag among multiple traffic flows will cause cell interleaving, thereby precluding packet reassembly at the egress ATM router. These limitations may present scalability problems in a large Internet environment.

The TAG Switching model requires that ATM switches support both standard ATM software and the TAG Switching software. This will allow a common ATM core to support TAG Switching as well as standard ATM switching (Figure 23).

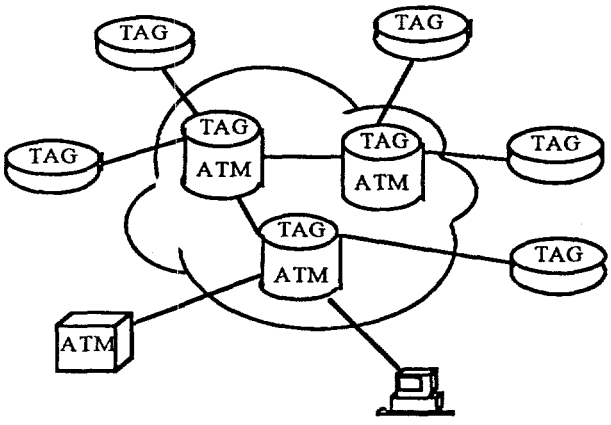


Figure 23. Common ATM core for TAG and ATM switching.

Cisco is attempting to form a TAG Switching working group at the IETF organization.

6. Summary

Both the IETF and the ATM Forum have proposed an overlay architecture to support IP over ATM. This overlay architecture requires duplicate address spaces as well as routing protocols which hide ATM's physical topology from the upper layer, IP. Therefore, this architecture increases the network management overhead as well as the network management complexity. This architecture models the ATM infrastructure as a logical Non-Broadcast Multi-Access (NBMA) medium. Using a server-based mechanism and complex protocols, the NMBA model emulates functions of the broadcast and multicast capabilities inherent in shared medium architectures. In addition to implementation complexities, servers may become performance bottlenecks and single points of failure. Nevertheless, the NBMA model will provide support for virtual local area networks (VLANs) across the ATM infrastructure independent of their physical location.

Many ATM vendors have implemented the IETF's CLIP model and the ATM Forum's LANE model. However, both models lack a mechanism to take advantage of ATM's cut-through paths, and therefore will experience performance bottlenecks at intermediate routers. Furthermore, neither supports the QoS capabilities promised by ATM. Moreover, the IP multicast functionality frequently found in integrated services applications is lacking in the IETF's CLIP model. In order to overcome these limitations, both the IETF and the ATM Forum are working on extensions to their current architectures. Ongoing work includes the IETF's MARS and NHRP, as well as the ATM Forum's MPOA.

Two proprietary IP-over-ATM solutions, IPSILON Networks' IP Switching and Cisco Systems' TAG Switching, have been proposed as simpler alternatives to the overlay architecture. These proposals model the ATM infrastructure as point-to-point networks which are, therefore, suitable for use as backbone solutions. Since today's protocols already work well with point-to-point networks, they will continue to work over these proprietary solutions without any modifications. Therefore, there is no need for additional complicated protocols and servers to emulate shared-medium broadcast and multicast. Moreover, by running standard routing packages in their ATM switches, this architecture preserves IP's connectionless paradigm while potentially achieving ATM's hardware speed, as well as eliminating the need for ATM signaling and routing. However, both models may have scalability problems in large Internet environments.

Because IPSILON Networks' IP Switching is based on IP flows, the number of entries in ATM switches may become large enough to present scalability problems in Internet environments. If there were shortages of switch table entries, IP flows would take the default hop-by-hop route without the benefit of short-cut performance. Today's IP Switching product supports a priority-based QoS service. In future releases, IPSILON plans to implement dynamic resource reservation using RSVP.

Cisco's TAG Switching is network topology driven and may, therefore, be as scaleable as today's routing protocols. However, when implemented in an ATM infrastructure, its scalability is limited by the following: 1) the ATM header lacks the necessary format to support more than two levels of IP's routing hierarchy, and 2) ATM TAG Switching cannot perform route aggregation, i.e., a shared tag among multiple traffic flows will cause cell interleaving. These limitations may present a problem in large Internet environments. The TAG Switching architecture is still under development. Currently there is no product available for proof of concept.

Acknowledgment

The authors appreciate the technical review and the useful suggestions from Don Hall and Christine Yang. We are especially grateful to Anthony Alles, Steve Deering, and Lixia Zhang for providing reference material from which we drew upon to develop this survey.

References

[1] M. de Prycker, "Asynchronous Transfer Mode - solution for broadband ISDN", Ellis Horwood Limited, ISBN 0-13-053513-3, 1991.

[2] R. Braden. D. Clark, S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June, 1994.

[3] ISI, USC, "DOD Standard Internet Protocol", RFC 760, January, 1980.

[4] ISI, USC, "Transmission Control Protocol", RFC 793, September, 1981.

[5] R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 1883, December, 1995.

[6] R. Braden, L. Zhang, , S. Berson, , S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP), Version 1 Functional Specification", Internet Draft, November, 1996.

[7] M. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, "RTP, A Transport Protocol for Real-Time Applications", January, 1996.

[8] W. Fenner, "Internet Group Management Protocol, Version 2", Internet Draft, October, 1996.

[9] T. Pusateri, "Distance Vector Multicast Routing Protocol", Internet Draft, September, 1996.

[10] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification", Internet Draft, November, 1996.

[11] A. Ballardie, S. Reeve, N. Jain, "Core Based Trees (CBT) Multicast- Protocol Specification", Internet Draft, September, 1996.

[12] Y.. Dalal, and R. Metcalfe, "Reverse Path Forwarding of Broadcast Packets", Communications of the ACM, Vol. 21 Num. 12, pp 1040-1048.

[13] C. Hedrick, "Routing Information Protocol", RFC 1058, June, 1988.

[14] The ATM Technical Committee, "ATM User Network Interface (UNI) Specification Version 3.1", Prentice Hall,1995, pp151-294.

[15] The ATM Technical Committee, "Private Network-Network Interface Specification Version 1.0", af-pnni-0055.000, March, 1996.

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

[17] J. Luciani, D. Katz, D. Piscitello, B. Cole, "NBMA Next Hop Resolution Protocol", Internet draft, October 1996.

[18] G. Armitage,, "Support for Multicast over UNI3.0/3.1 based ATM Networks",RFC 2022, November, 1996.

[19] The ATM Forum Technical Committee, "LAN Emulation over ATM Specification-Version 1", ATM Specification, February, 1995

[20] The ATM Forum Technical Committee. "Multi-Protocol Over ATM Version 1", ATM Forum/BTD-MPOA-MPOA-01.11 Draft, February, 1997

[21] P. Newman, W. Edwards, R. Hinden, E. Hoffman, F. Ching Liaw, T. Lyon, G. Minshall, "Transmission of Flow Labelled IPv4 on ATM Links Ipsilon Version 1.0", RFC 1954, May, 1996

[22] Y. Rekhter, B. Davie, D. Katz, E. Rosen, "Tag Switching Architecture - Overview", Internet Draft, January, 1997.

[23] V. Fuller, T. Li, J. Yu, K. Varadhan, "Supernetting, an Address Assignment and Aggregation Strategy", RFC 1338, June, 1992.

[24] J. Moy, "The OSPF Specification", RFC 1131, January 1993.

UNLIMITED RELEASE

INITIAL DISTRIBUTION:

1	MS 0451	T. D. Tarman, 9417
1	MS 0806	M. O. Vahle, 4616
1	MS 0806	S. A. Gossage, 4616
1	MS 0806	L. Stans, 4615
1	MS 1436	LDRD Office
1	MS 9003	D. L. Crawford, 8900
1	MS 9011	R. E. Palmer, 8901
1	MS 9011	P. W. Dean, 8910
1	MS 9011	J. M. Brandt 8910
10	MS 9011	H. Y. Chen, 8910
1	MS 9011	J. A. Hutchins, 8910
1	MS 9011	R. Tsang, 8910
1	MS 9011	J. Meza, 8950
1	MS 9011	D. B. Hall, 8970
1	MS 9012	J. E. Costa, 8920
1	MS 9012	S. C. Gray, 8930
1	MS 9019	B. A. Maxwell, 8940
1	MS 9001	T. O. Hunter, 8000; Attn: J. B. Wright, 2200 J. F. Ney (A), 5200 M. E. John, 8100 L. A. West, 8200 W. J. McLean, 8300 R. C. Wayne, 8400 P. N. Smith, 8500 T. M. Dyer, 8700 P. E. Brewer, 8800
3	MS 9018	Central Technical Files, 8940-2
4	MS 0899	Technical Library, 4916
1	MS 9021	Technical Communications Dept., 8815/Technical Library, 4916
2	MS 9021	Technical Communications Dept., 8815 For DOE/OSTI