EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE LAUSANNE
POLITECNICO FEDERALE DI LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY LAUSANNE

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Communication Systems Division (SSC)
EPFL CH-1015 Lausanne, Switzerland
http://sscwww.epfl.ch

IP and ATM - a position paper

Silvia Giordano Rolf Schmid Reto Beeler
Hannu Flinck Jean-Yves Le Boudec

July 1997

Technical Report SSC/1997/017

IP and ATM - a position paper

Silvia Giordano, LRC-EPFL; Rolf M. Schmid, Ascom Tech Ltd. (Editor); Reto Beeler, Ascom Tech Ltd. ; Hannu Flinck, Nokia Research; Jean-Yves Le Boudec, LRC-EPFL

Abstract:

This paper gives a technical overview of different networking technologies, such as the Internet, ATM. It describes different approaches of how to run IP on top of an ATM network, and assesses their potential to be used as an integrated services network.

Keyword List: ATM, IPv4, IPv6, RSVP, LANE, CLIP, NHRP, MPOA, Arequipa, IP-Switching, Tag Switching

Note: this document was produced in the framework of the ACTS project EXPERT.

Executive Summary

For many years, the ATM based Broadband-ISDN has generally been regarded as the ultimate networking technology, which can integrate voice, data, and video services. With the recent tremendous growth of the Internet and the reluctant deployment of public ATM networks, the future development of ATM seems to be less clear than it used to be.

Given the vast installed base of IP networks and equipment today together with the unmatched variety of IP based applications, the key to the success of ATM in the short and medium term will be its ability to allow for interoperability between itself and these technologies.

This paper gives a technical overview on the competing integrated services network solutions, such as IP, ATM and the different available and emerging technologies on how to run IP over ATM, and tries to identify their potential and shortcomings.

Table of Contents

1. Introduction

For many years, ATM based Broadband-ISDN has generally been regarded as the ultimate networking technology, which can integrate voice, data, and video services and which is suitable for LANs and WANs, both private and public.

While ATM has been around for quite a while now the initially expected fast take-off of ATM in the public WAN area did not take place. Although ATM products are broadly available today, public network operators hesitate with the deployment of public ATM based networks. In contrast to this, ATM had quite an impact in the private LAN area, where ATM is mainly deployed as a highspeed backbone network interconnecting legacy LAN equipment, driven by the need to increase transmission speed.

With the recent tremendous growth of the Internet, the future role of ATM seems to be less clear than it used to be. WWW based use of multimedia applications on the Internet is widespread. By offering not only typical data services but even real time voice and video applications (though with poor quality), the Internet is entering the typical target market of ATM at service level. Furthermore the Internet Society is quite drastically loosening their policy of shared resources and free usage and start investigating on how to introduce resource reservation and charging support in the Internet to provide better support for multimedia applications and service providers.

The dominance of IP based networks in the WAN and LAN area has also led to proposals for ATM deployment that considerably differ from the traditional view of public telecom operators, such as using ATM only as a high speed transmission system.

The discussions about whether IP or ATM is the better technology for an integrated services network are ongoing and reached almost the state of a ‘war’ between advocates of the two technologies.

This position paper written by the ACTS-EXPERT projects gives a technical overview of the different technologies today and makes a neutral assessment of their feasibility for an integrated services network. In order to be able to compare different solutions, we establish the requirements for an integrated services network in section 2. These requirements depend on the perspectives of different actors and contain more than just the requirement for quality of service (QoS). Then we focus on the main competitors, the Internet protocols (section 3) and the ATM technology (section 4). In section 5 we discuss several of today’s and tomorrow’s solutions for IP support on top of ATM networks. Of each presented technology the advantages and disadvantages are assessed in the corresponding section. Section 6 compares the technologies and tries to guess about their applicability and the role they are going to play in the future.

2. Integrated Services Networking Requirements

In this section, some of the most important requirements are listed for the comparison of integrated services networking technologies. Criteria for a broad acceptance of such a technology are established. The requirements can be categorised as follows:

- Requirements from the *user*'s perspective
- Requirements from the *service provider*'s perspective
- Requirements from the *network provider*'s perspective

2.1 User's Perspective

Because the acceptance and success of a technology is dependent on user requirements, it is essential that these requirements are met by the technology. A user's major concerns are performance, ease of use, cost, universal availability and security of services:

- **Guaranteed support of an appropriate minimal performance**: Each service should be offered with the *appropriate* minimal performance. It is important to note here that from the user's perspective not all services have to be offered with very high performance. However there are considerable differences in performance requirements depending on the kind of service, and on its pricing. An audio phone service with today's POTS performance imposes high quality requirements on a networking technology, whereas an e-mail service could be acceptable with as basic a quality requirement as reliable delivery. In order to be generally acceptable to the user, services using a new networking technology should be offered with equal or better performance and at the same or even lower price than those commonly available today. A **Cost-Performance compromise** has to be taken into account and it can be argued that the final decision for such a compromise ideally should be delegated to the user. This requires some degree of **Cost-Performance transparency**.
- **Ease of Use** (and configuration) is important and the availability of **simple and cheap terminal equipment** is a must. This naturally calls for **Integration of services**: An acceptable networking technology should integrate the full range of services to satisfy all of today's and tomorrow's communication needs.
- **Universal connectivity** is a growing requirement. Users would like to have the possibility of reaching any other user over the same access technology. Flexibility features such as personal mobility should be supported as users wish to have access to their subscribed services from any terminal equipment.
- **Security**: An acceptable networking technology should support security features such as authentication and privacy. Authentication limits service access to authorised users only, e.g. eliminating the risk of users accessing a service without paying for it. Privacy means encryption of data so that eavesdropper are not able to interpret the received data, allowing for example credit card numbers to be exchanged over the network.

2.2 Service Provider's Perspective

With deregulation in the public telecommunication market and the transition towards integrated services networks, it is expected that there will be a clear functional separation between network providers, who operate network infrastructure and provide network connectivity, and service providers, who provide services on top of that network infrastructure. In today's public networks, both network provisioning and service provisioning is typically under the control of the same legal entity (e.g. PNO), but in the future we expect a high number of independent service providers to enter the market, who will run their business separately from network providers [1].
The requirements for an integrated services networking technology are not identical for service providers and network providers. This section lists the requirements from the service provider's perspective. Service provider's requirements also include requirements imposed by the provided services themselves. It has to be noted here that a "service" in this context is not restricted to new multimedia services.

Service provider requirements are mainly in the area of universal connectivity and traffic support, network and service separation support, service management and security:

- **Universal connectivity** and **universal traffic support** are essential to maximise the market size for a service. An integrated services networking technology should be able to cope with any traffic type, as traffic characteristics produced from different services vary considerably. This calls for **high bandwidth availability** so that service providers are able to offer any kind of service. It also requires **end-to-end transmission quality guarantee** in order to be able to provide services with the user-requested performance. **Addressing flexibility** is needed as many services require more than just normal unicast (point-to-point) addressing. The networking technology should offer the flexibility to support the full range of addressing types such as unicast, multicast, broadcast and anycast.
- **Service charging support**: In an integrated services network there should be support for service charging as service providers are expected to prefer to charge for their services independent from network providers.
- **Security**: In addition to the security related requirements from the user's perspective (i.e. authentication and privacy), service providers require support of non-repudiation. Non-repudiation means that once a user has committed to pay for a service, the payment can not be refused.
- **Network and service separation support** and **ease of service management** are not taken into account in this paper as these issues are not primarily dependent on the underlying network technology.

2.3 Network Provider's Perspective

Even though deregulation in the telecommunication market will allow for new network providers, the group of network providers is the smallest group of actors in an integrated services network. Nevertheless their requirements must not be neglected because they build, operate and own the networks.
The main focus of network providers is on manageability, network availability, scalability, chargeability and (of course) low costs:

- **Scalability**: A «future-proof» networking technology should be able to scale to an unlimited number of endpoints and to ever increasing resource demands.
- **Support of network charging**: Charging of network usage is an essential requirement of network providers. In integrated services networks it will not be sufficient to have a flat-rate usage charging but rather a usage charging based on traffic size and quality. Without this traffic based charging, network overload situations will become the norm. Traffic based charging can only be achieved if a networking technology provides the functionality to monitor the traffic.
- **Low cost**: The infrastructure as well as the operating costs for a networking technology should be low. Protection of investment is an important factor. Given the enormous investments in existing networking infrastructure (e.g. POTS, Cable TV), the ability to be run on top of parts of this existing infrastructure is very essential for a new networking technology in the opinion of network operators. This holds especially true in the customer access area where investments for physical connections are huge. Furthermore a new networking technology should allow for a smooth migration, making use of large parts of existing telecommunication infrastructure for the short term and allow for successive replacement with new infrastructure for the medium and long term.
- **Network management support** and **network availability** are not taken into account in this paper, as these issues are not primarily dependent on the underlying network technology.

3. Internet Technology

3.1 IPv4

3.1.1 General Overview

The Internet Protocol (IP, or IPv4) is the central part of the Internet protocol suite. IP (RFC 791 [2], RFC 1122 [3]) offers a *connectionless* packet delivery service on top of which the transport level protocols i.e. TCP and UDP build their functionality. IP is a *datagram oriented* protocol that treats each packet independently. Therefore each packet must contain complete addressing information. It neither guarantees delivery nor integrity, because the protocol does not use checksums to protect the content of the packet and there is no acknowledgement mechanism to determine whether the packet has reached its destination or not.

The IP protocol together with a set of supporting protocols (ARP, RARP or BootP, ICMP) defines the format of the Internet datagram, addressing, address resolution, packet processing, routing, and error reporting mechanisms. As described in RFC 1122 [38] any host running the IP protocol suite typically also supports the following protocols: Address Resolution Protocol (ARP, RFC 826[4]) and Internet Control Message Protocol (ICMP, RFC 1122 [3]). The following figure summarizes how the IP protocol is related to the other protocols in the Internet stack. As shown in the figure, IP can be run over a variety of data link layers, because IP *hides* the underlying technologies from its users.

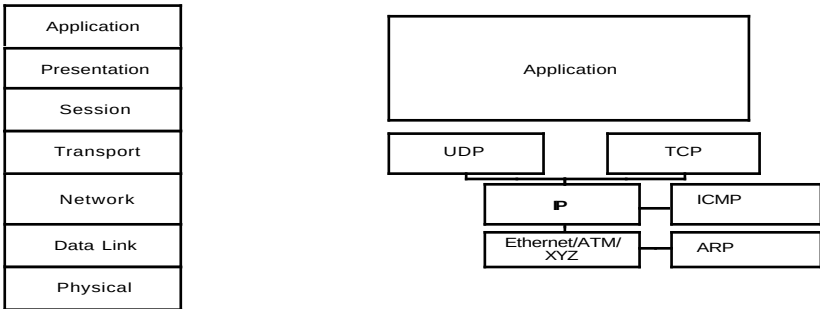

Figure 1: IP in the Protocol Stack

3.1.2 The IP datagram structure

The IPv4 datagram is variable in length with a theoretical maximum of 65’535 octets. However, in practice the size of a datagram is limited by the size of the data link layer or the physical layer as a whole datagram has to fit into a single frame of the underlying layer. For example, Ethernet limits the datagram sizes to 1500 octets. This limitation to the datagram size imposed by the underlying technology is called the “maximum transfer unit”, MTU. However, in a heterogeneous environment with varying MTUs, the datagram may need to be fragmented into smaller pieces, IP therefore supports fragmentation.

The coding of an IP datagram format is shown in the following figure.

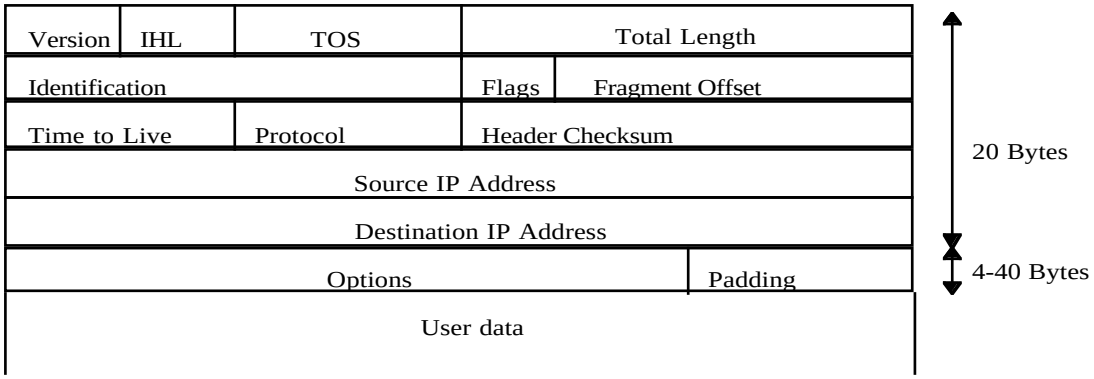

Figure 2: IP datagram format

- Version: identifies the protocol version (i.e. 4 for IPv4)
- Internet Header Length (IHL): the length of the header in 32 bit words
- Type of Service (TOS): indicates possible priority and the type of transport the datagram desires (options are low delay, throughput, reliability)
- Total Length: the length of the datagram measured in octets up to 65'535 octets
- Identification: a value assigned by the sender to aid in assembling the fragments of a datagram
- Flag bits: control the fragmentation (e.g. don't fragment, may fragment,...)
- Fragment offset: indicates the place in the original datagram where this fragment belongs
- Time to Live: indicates the maximum hop number that the datagram is allowed to pass in the network
- Protocol: indicates the next layer protocol that was used to create the user data (e.g. TCP, UDP)
- Header Checksum: a 16 bit checksum over the header
- Source/Destination address: 32 bit IP addresses to identify the sender and receiver
- Options field: carry information for network control, debugging, routing and measurements.

3.1.3 IP Addressing and Routing

In IPv4 the IP address space is limited to 32 bits. An address begins with a network number used for routing, followed by a local, network internal address. IP addresses are classified in four classes according to the size of the network portion of the address:

- *Class A*, where the high order bit is zero, the next 7 bits are for the network, and the rest for the local address
- *Class B*, the high order two bits are one-zero, the next 14 bits are for the network and the rest for the local address
- *Class C*, the high order three bits are one-one-zero, the next 21 bits are for the network and the last 8 for the local address.
- *Class D*, this is for multicasting, the high order four bits are one-one-one-zero followed by multicasting address

As can be seen from the above address classes IP supports multicasting (in subnetworks). Broadcasting is also supported (RFC 919 [5]).

The IP addressing builds upon the notion of "network", which is fundamental for routing in the Internet. There are two types of equipment at the IP level, hosts and routers. Hosts are any end-user computer system that connect to a network. Hosts know (or learn during the boot phase) their network address and local address. This forms the host address. A physical host may have several local addresses and a single network address. A multi-homed host is a host that is attached to two different networks as a host, however this is a special case. A router is a (dedicated) computer that attaches to two or more networks and forwards packets from one to the other based on the network portion of the destination IP address. Routers exchange network addresses as reachability information between them using various routing protocols (e.g. EGP, OSPF) depending on where in the network hierarchy the routers are located.

IP traffic within the same network can be delivered directly from host to host, whereas IP traffic to another network always passes one or several routers.

3.1.4 Assessment

The roots of IP are in the early '80s. Since then processing power and memory size of computers and the nature of applications have changed considerably. IPv4 has the following restrictions, some of which have led to the recent redesign of the IP protocol (IPv6, see section 3.2):

- The fixed size address space of 32 bits is a limiting factor for the predicted Internet growth (B class addresses exhausted, supernetting of C class addresses is only a short term solution).
- New types of address hierarchy are needed to make the protocol more flexible.
- No support of an anycast addressing concept.
- Per packet computational load is not optimal and can be improved, resulting in a more efficient datagram delivery.
- No support of multimedia type of streaming (flows).
- No support of guaranteed QoS: just plain, best effort, connectionless packet delivery.
- Potentially inefficient routing (all IP packets of a persistent data flow are routed independently)
- Potentially inefficient transmission (IP header is too big, especially for short packages)

If only the mandatory set of the IPv4 protocol suite (as described in [38]) is supported, the following restrictions also apply:

- No plug and play type of address autoconfiguration and re-numbering.
- No network layer security support.
- No mobility support.

However, there are additional RFCs which cover these features.

Despite these restrictions IP is widely deployed today and with the current boom of the Internet it will become even more important among the network layer protocols. Apart from its wide deployment offering almost universal connectivity, there are some other advantages of IP:

- There is an unmatched variety of services and applications available that build on IP
- IP can be run over a big variety of physical layers
- IP is a working solution and its performance has been well tuned over the years
- IP equipment is cheap for network providers as well as for users
- IP based applications do not have to know their bandwidth demand in advance and can easily adapt to the encountered traffic level along the traversed network path.
- Separating network and service provisioning is a reality in the Internet architecture

3.2 IPv6

3.2.1 Why IPv6

Mainly triggered by the fear of the approaching address space exhaustion and to solve some of the shortcomings of IPv4 (see section 3.1), the IETF started working on IPv6 (or IPng) in 1992. By 1996 version 6 of the Internet Protocol was specified.

IPv6 is not a radical change to IPv4, it is rather an evolutionary step and coexistence between Ipv4 and Ipv6 is possible for a transition phase [11]. Except for the larger address space and some autoconfiguration features, all new functionality could also have been fitted into IPv4. Nevertheless, after over 10 years of building and enhancing the Internet protocol stack, it is necessary to clean and consolidate the functionality of the very central IP layer and make it a ready platform which will be able to cope with new Internet functionality required in the near future.

3.2.2 The IPv6 protocol suite

The IPv6 protocol suite is not defined in a single specification but comes in a whole collection of RFCs, the most important of which are listed below:

- IPv6 (RFC1883 [6]), IPv6 Addressing (RFC 1884 [7])
- ICMPv6 (RFC 1885 [8]) Internet Control Message Protocol, including address resolution
- Authentication Extension (RFC 1826 [9]), ESP Extension (RFC 1827 [10])

All higher layer protocols in hosts (UDP, TCP, Web, DNS, ..) need to be enhanced to be able to use the new functionality of IPv6. There are Internet drafts available on how to enhance the IPv4 API (socket interface) in order to bring the IPv6 functionality to the application layer.

3.2.3 New addressing and routing

IPv6 uses 16 byte addresses and improves addressing flexibility through the definition of unicast, multicast and anycast addresses. Furthermore IPv6 supports plug and play features such as automatic IP address configuration and re-numbering.

Scalability was introduced to IPv6 multicast routing by using address scopes. In general, IPv6 routing is almost identical to CIDR of IPv4, based on the route selection of longest matching address prefix. With very little modification, all of IPv4's routing mechanisms can be used to route IPv6.

Source routing is used in IPv6 to ease future implementation of new functionality such as terminal mobility and provider selection.

3.2.4 IPv6 packet structure

The IPv6 packet's *base* header is a streamlined IPv4 header, reducing the processing cost of packet handling and limiting the packet size by removing some of the fields and options. Some of the options removed from the old header and some new options of IPv6 are now supported through an arbitrary number of *extension* headers following the base header, each of them indicating in its Next Header field the type of the next following extension header. Examples for such extension headers are the source routing extension header, the fragmentation header and the authentication header. Extension headers are normally only examined or processed by the destination node. The use of extension headers introduces high modularity in the IP packets and easily allows future options or extensions to be integrated. The data follows the last extension header. The structure of the IPv6 base header and of an IPv6 packet is given in Figure 3.

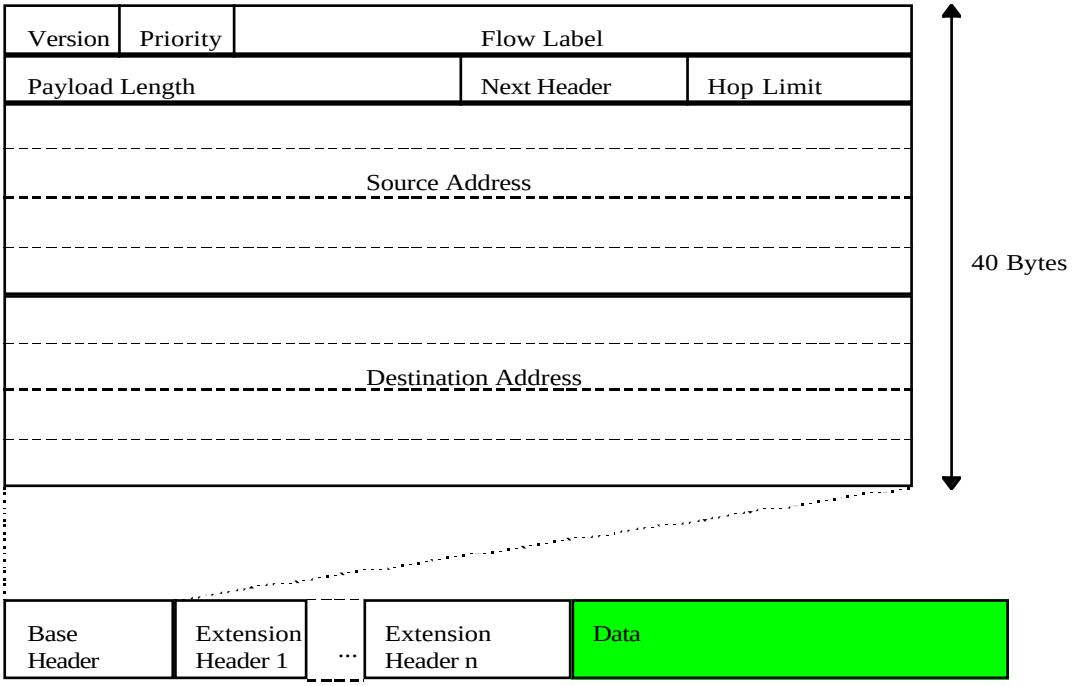

Figure 3: the IPv6 packet structure

3.2.5 New features of IPv6

IPv6 introduces **flow labelling** capabilities (Flow Label field in base header), which allows packets belonging to the same flow to be labelled. The sender can request special handling of a flow by the routers,

such as non-default quality of service or “real-time” service. Routers can cache flow information obtained from processing the first packet of a flow and thus speed up processing for following packets of the same flow.

The Priority field in the base header allows the desired **delivery priority** of a packet to be specified, but, this is only relative to the priority of other packets from the same source.

IPv6 introduces **network layer security** defined in the authentication and the ESP extension header. Authentication is used to guarantee the packet sender’s identity and ESP means the encapsulation of security payload so that a third party can not read it.

IPv6 supports **source routing** capabilities defined in another extension header.

IPv6 allows only **restricted fragmentation,** i.e. fragmentation is only allowed at the packet source but NOT at intermediate IPv6 routers. This is achieved by either using the minimum MTU guaranteed by all delivery systems (576 octets) or by using ICMPv6 messages for path MTU discovery.

3.2.6 Transition from IPv4 to IPv6

IPv6 and IPv4 are similar but all the same are distinct protocols. To allow for an incremental upgrade of IPv4 equipment to IPv6, during which both protocols can coexist, it is crucial that there is both a way to interwork between the two protocols and to tunnel IPv6 through a cloud of IPv4. Given today’s vast installed base of IPv4 equipment this was a key issue during IPv6 protocol design.

To address the transition problem, there is RFC 1933[11], which defines a set of mechanisms that IPv6 hosts and routers should implement in order to be compatible with IPv4 hosts and routers. The proposed solution is based on dual IP layer nodes, tunnelling, and DNS support.

- Dual IP layer node is a host or router that implements both IPv6 and IPv4
  All such nodes need an IPv6 and an IPv4 address. For this purpose the “IPv4 compatible IPv6 address” was defined, which uses the IPv4 address in its lower 4 bytes and all zeros for the 12 higher bytes.
- Tunnelling defines how IPv6 packets have to be encapsulated within IPv4 datagrams, so that they can be carried over IPv4 routing infrastructure, and addresses the tricky issues of fragmentation and ICMP error message mapping.
- DNS is used to provide IPv4 and IPv6 addresses for a machine name.

It is very important to note here that this is only a short term solution which will work as long as the IPv4 address space is not exhausted, as all dual layer nodes need IPv6 and IPv4. Other, much more complex solutions, such as real address translations, will have to be defined after the address space exhaustion.

3.2.7 State and availability of IPv6

Work on the IPv6 protocol has been finished though there is still some remaining work on higher and lower layers, such as to support IPv6 in all routing protocols over all different media and enhance the API for IPv6.

There is already a variety of router and host implementations available from different vendors and for different architectures. A global IPv6 based backbone (6bone) is operational and growing.

3.2.8 Assessment

IPv6 is a neat new version of IP, cleaning up old functionality and adding some new functionality to make it a stable and future proof network layer. It comes as a SW upgrade to current IPv4 equipment and offers an incremental transition phase, minimizing costs for new equipment and protecting past investments.

A very important new feature of IPv6 is its basic support of QoS at the network layer through flow labels and priority indication. This does not mean however, that IPv6 alone can guarantee real end-to-end quality of service as there is no way to make network resource reservations. But IPv6 provides the network layer functionality which allows end-to-end quality of service to be provided when used together with protocols for network resource reservation like RSVP (see section 3.3). IPv6 basic QoS support also provides the basic functionality for a future, traffic based charging.

The larger address space overcomes the current limits of the Internet growth and has the potential to provide world-wide, universal connectivity. The big challenge for IPv6 is for its transition to be completed before IPv4 routing and addressing break. If this can not be achieved, very complex address translation solutions would have to be used to be able to keep the Internet paradigm of universal connectivity alive. This puts a lot of pressure on finishing standardisation work and speeding up the deployment of IPv6.

The security features of IPv6 support authentication and privacy. They also provides the basic functionality for service charging.

With its new plug and play features, IPv6 networks are much easier to configure and maintain.

All higher layer protocols and applications need to be ported to be able to make use of the new functionality provided by IPv6.

3.3 Resource ReSerVation Protocol (RSVP)

3.3.1 Motivation

IP provides best effort datagram delivery that is sufficient for most of the conventional applications such as e-mail, WWW and file transfer. However, a new class of application (e.g. multimedia) is emerging that requires guaranteed resources from the network in order to function properly. Typically such requirements for resource guarantees are related to stringent real time requirements.

To address the problem of resource reservation in the Internet, the IETF formed the Integrated Services working group. This working group with the goal of "efficient Internet support for applications that require service guarantees" is defining an Integrated Services framework, of which RSVP is an integral part.

It has to be noted here that RSVP is enhancing IP based networks to support end-to-end quality of service, it is however not related to ATM.

RSVP is a signalling protocol for the Internet.

3.3.2 Protocol overview

Resource ReSerVation Protocol, RSVP [12][13], has been proposed to be the protocol that allows applications to reserve network resources in an IP network such as the Internet. It is currently in draft state in the IETF. RSVP operates on top of IP (either IPv4 or IPv6) and it relies on standard Internet routing. It is used both in hosts and routers to reserve resources for a simplex (uni-directional) data stream, called a *flow*. A flow is a sequence of datagrams identified either by the IP destination address (either multicast or unicast address), or by the IP protocol ID and optionally by a destination port. The requested QoS for the flow is described by a *flowspec* together with a *filter spec*. These two form a *flow descriptor* that is carried in the resource reservation message.

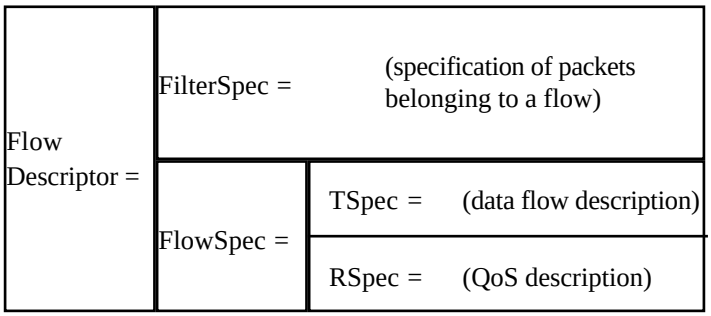

Figure 4: RSVP flow descriptor

RSVP is designed for both unicast and multicast communication in a heterogeneous network, where receivers may have different characteristics and multicast membership is dynamic. These requirements lead to a solution, where the *receiver* is responsible for initiating the resource reservation.

The message flow for the establishment of a network reservation for a multicast communication is shown in Figure 5.

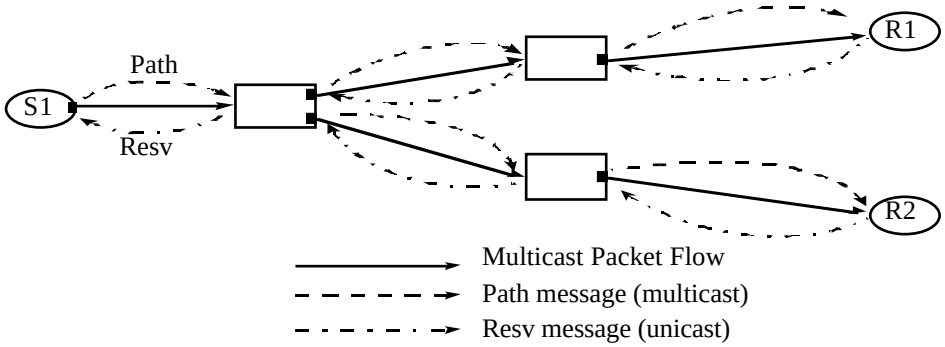

Figure 5: RSVP message exchange in the multicast tree

It is assumed that a multicast group already exists (created by Internet Group Management Protocol, IGMP [14]). The sender S1 sends a Path message to a multicast group announcing the characteristics of the flow it is going to send. The Path message contains a Tspec, describing the maximum traffic characteristics of its data flow, and a Filter Spec, describing the packet format of the flow. When the receivers, R1 and R2, want to make a resource reservation, they will send a Resv message upstream following exactly the inverse path of the Path message. The Resv message contains the desired reservation style (see Figure 7) and flow descriptor. The Resv message creates reservation state in each RSVP capable router along the path from the receiver to the sender. In a multicast situation as the one shown in Figure 5 there are nodes that will receive two or more Resv messages from different branches of a multipoint tree. These nodes merge the received reservations and forward only one merged reservation request upstream, containing the most demanding (maximum) flowspec.

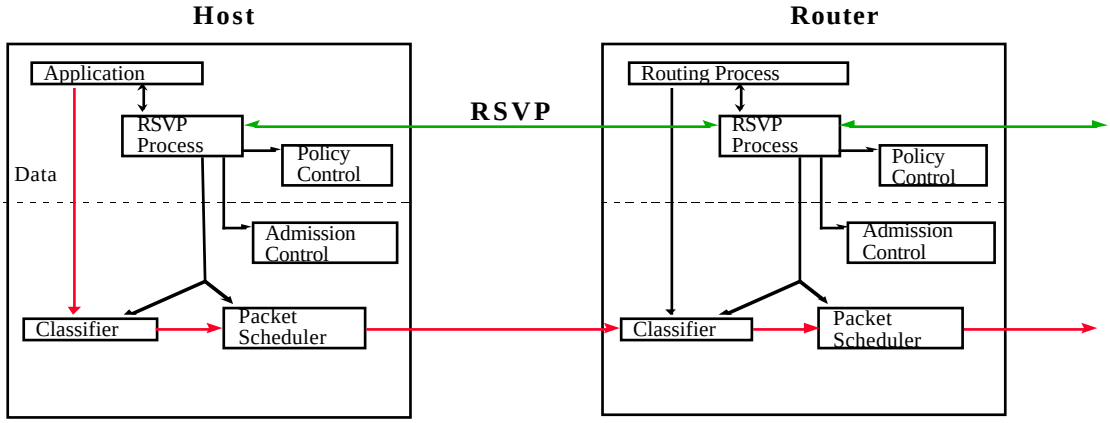

Figure 6: RSVP in hosts and routers

The resource reservation request indicated in the Resv message has to pass admission control and policy control modules in all RSVP equipped routers and hosts on its way. These check if the reservation can be accepted. Admission control determines whether the node has sufficient resources and policy control [15] deals with administrative issues such as accounting and access rights. If the reservation passes these two checks, flow related parameters are set in the packet classifier and packet scheduler. If either of the checks fail, an error notification is returned. The packet scheduler is responsible for negotiation with the link layer to reserve the transmission resources. It is here that mapping from the flow level QoS to the link layer QoS takes place.

The treatment of RSVP reservations in routers depends on the reservation style indicated by the Resv message. The different styles and attributes are listed in the following figure.

Sender Selection	Reservations: Distinct	Reservations: Shared
Explicit	Fixed-Filter FF(S1{Q1}, S2{Q2}, ...)	Shared-Explicit SE((S1, S2, ...){Q})
Wildcard		Wildcard-Filter WF(*{Q})

Figure 7: RSVP reservation attributes and styles

The reservation styles indicate either if there should be a separate reservation for each sender of a session (Fixed-Filter), or if the reservation can be shared among the named senders of the session (Shared-Explicit), or if the reservation can be shared by all the senders (Wildcard-Filter). Fixed-Filter, Shared-Explicit and Wildcard-Filter style are mutually incompatible. This results in rules for merging the reservations. For example, merging of shared reservations with distinct reservations is prohibited.

RSVP uses *soft state* for the reservation. This means when a reservation is made, it must be periodically refreshed (suggested refresh period is currently 30 seconds). Refreshing is accomplished by sending Path and Resv messages. The advantage of using soft state for the reservation is that the route of the connection can be changed dynamically inside the network and the reservation will be re-established when the new Path and Resv messages has passed through the new route. Soft states also help to allow for dynamic multicast group membership

In addition to Resv and Path messages RSVP has messages for tearing down the reservation state. The PathTear message is sent from the sender to tear down the path and thus the reservation state and ResvTear is sent from the receiver. A sender can request reservation confirmation to its Resv message, the sender or a router that is merging the reservation to another reservation sends a ResvConf message to confirm the reservation.

3.3.3 Assessment

RSVP defines an efficient, flexible and robust solution for setting up resource reservations in IP based networks, but it does not scale well for large number of flows. RSVP is especially tailored for the need of multicast connections in heterogeneous networks.

With the support of resource reservation in the network application requested end-to-end QoS becomes possible. However it remains unclear, how routers are going to map the resource reservations to internal settings for the packet classifier and scheduler and how reliably they are going to support the requested reservation. Furthermore end-to-end QoS can only be guaranteed if all the routers and hosts along the routed path are running RSVP software, because tunnelling through non-RSVP clouds destroys all end-to-end QoS.

The fact that RSVP sets up reservations in the upstream direction of a pre-established multicast tree makes it impossible that QoS information is used for routing decisions.

The use of RSVP in the Internet may provide input for traffic based charging.

There is an IETF RSVP Working Group that is in charge of evolving the RSVP specification. The RSVP-WG also coordinates its work with the parallel IETF working group that is considering the service model for integrated service, in order to have RSVP compliant with the overall integrated service architecture and the requirements of real-time applications. The RSVP-WG also coordinates its work with the IPng-related working groups.

4. ATM Technology

4.1 Introduction

Asynchronous Transfer Mode (ATM) is a *cell-based, connection-oriented* switching technology that is designed to support a wide variety of services, including cell relay, frame relay, SMDS, and circuit emulation. ATM transmits all information using small (53 byte) fixed length cells over broadband or narrowband transmission facilities. It is *asynchronous* because the cells carrying user data are not required to be periodic. The asynchronous and multimedia characteristics of ATM are what makes it possible for ATM networks to carry both circuit and packet types of traffic simultaneously, with complete transparency to the applications. ATM was designed to provide large amounts of bandwidth economically and on-demand. When a user does not need access to a network connection, the bandwidth is available for use by another connection that does need it.

The ATM technology was defined by the ITU-T, mainly formed by the representatives of public network operators. The rather slow development of standards in ITU-T was sped up by the ATM Forum (founded in 1991), a growing group of companies focusing on private network and data communication.

The term ATM can be interpreted in a variety of ways. In fact, it is true to say that there is no single definition. It takes on many forms, encompasses both hardware and software, and can run on several types of digital transmission facilities. ATM can refer to a physical interface (the 53-byte cell), a switching technology, or a unifying network technology that provides integrated access to multiple services.

There is a broad consensus that ATM will first be implemented within wide area networks primarily as a switching technology to support existing services in private WANs and in public service networks. ATM excels primarily as a backbone technology, because it is in this context that most of the benefits of cell relay are realised.

4.2 Virtual Paths and Virtual Channels

Each ATM cell contains a two-part address, a Virtual Path Identifier (VPI) and a Virtual Channel Identifier (VCI), in the cell header. This address uniquely identifies an ATM virtual connection on a physical interface. The physical transmission path (such as DS1 or DS3) contains one or more virtual paths, and each virtual path can contain one or more virtual channels.
The VPI and VCI are tied to an individual link on a specific transmission path, and have local significance only to each switch. The VPI and VCI addresses are translated at each ATM switch in the network connection route -- each switch maps an incoming VPI and VCI to an outgoing VPI and VCI. Therefore, these addresses can be reused in other parts of the network as long as care is taken to avoid conflicts. ATM can perform switching on a transmission path, a virtual path, or a virtual channel.

4.3 Permanent Virtual Circuits and Switched Virtual Circuits

ATM provides two virtual circuit communications services: Switched Virtual Circuits (SVCs) and Permanent Virtual Circuits (PVCs). SVCs establish short-term connections that require call setup and teardown, while PVCs are similar to dedicated private lines because the connection is set up on a permanent basis. Users establish PVCs either by requesting them from a public carrier providing the frame relay or ATM service, or from the WAN administrator of the private network. ATM virtual connections can operate at a constant bit rate (CBR) for voice and video traffic, at a variable bit rate (VBR) for bursty traffic and at available bit rate (ABR) or unspecified bit rate (UBR) for best effort traffic. Each virtual connection has its own set of parameters (Minimum Cell Rate (MCR), Sustained Cell Rate (SCR), Peak Cell Rate (PCR)), that determine the amount of bandwidth, priority, Quality of Service (QoS), etc.

4.4 ATM Signalling, Routing and Addressing

ATM Signalling Protocols vary by the type of ATM link - ATM UNI signalling is used between an ATM endsystem and an ATM switch across an ATM UNI; ATM NNI signalling is used across NNI links.

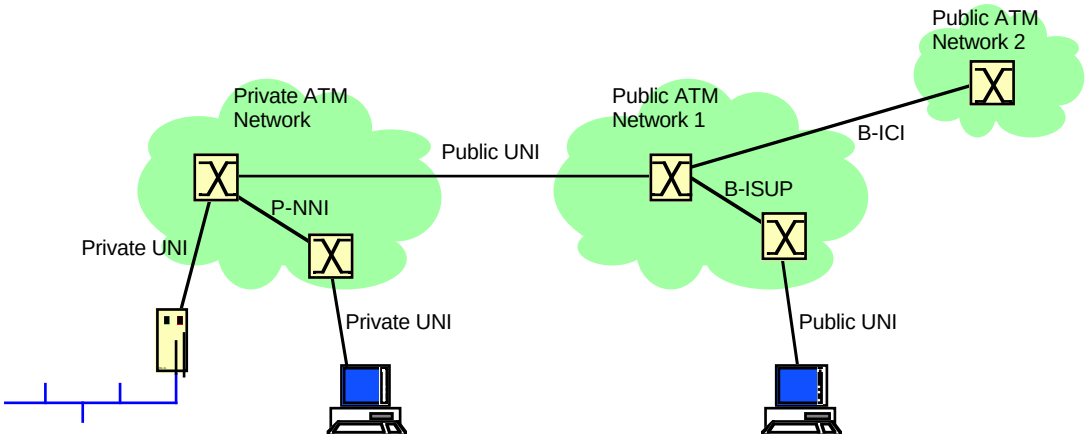

Figure 8: ATM network architecture and interfaces

The current standard for UNI signalling is described in the ATM Forum UNI4.0 specification [16], which is an enhancement to the earlier UNI3.1 and provides some kind of alignment to the recommendations for public UNIs specified by ITU-T. Besides the basic call set up and tear down functionality UNI4.0 defines point-to-multipoint operation, address registration and extended QoS support. An overview of UNI signalling capabilities can be found in [17].

ATM switches are interconnected via one of three NNIs: P-NNI in private ATM networks, B-ISUP in public networks and B-ICI between different public networks. NNI interfaces define not only signalling procedures but also routing. P-NNI defined by the ATM Forum [18] supports not only signalling procedures similar to UNI4.0 but also topology discovery via the distribution of reachability information, hierarchical routing and addressing and QoS.

Whereas ITU-T has long settled upon the use of telephone number-like E.164 addresses for public ATM networks, the ATM Forum defined a private address format based on the syntax of an OSI Network Service Access Point (NSAP) address to be used in private networks.

Application Programmer Interfaces (APIs) for ATM are still under definition. The industrial standard WinSock2 (for Windows based applications) is becoming available now.

4.5 Assessment

The most significant advantages of pure ATM solutions are:

- ATM supports end-to-end QoS guarantees on a per virtual connection basis. ATM virtual connections allow users to expect a guaranteed minimum amount of bandwidth for each connection. ATM supports several Quality of Service (QoS) classes to accommodate the differing delay and loss requirements for each type of traffic.
- ATM uses statistical multiplexing, which allows bandwidth to be shared among many users. Bandwidth is only provided when it is needed “on demand”, thus reducing the cost of network resources.
- ATM supports multiple services. ATM can be used to transport literally any kind of information and can simultaneously support a broad range of user interfaces. Only ATM WANs can provision frame relay, SMDS, native ATM, voice, video, and existing leased circuit services (circuit emulation) over the same wide area circuits.
- ATM provides high performance.
- ATM enables traffic based charging.

On the other hand, ATM has got some disadvantages that may interfere with widespread deployment, ease of high-speed implementation or present architectural concerns.

- ATM technology is very complicated and the control software is extensive and complex.
- Connection establishment times may be prohibitive for short duration data flows.
- Applications have to know their QoS demands in advance and can not easily adapt to changing network load.
- No multipoint-to-multipoint support

- Currently there are hardly any applications available that run directly on top of ATM and can exploit its benefits. ATM APIs are only emerging now and today's TCP/IP based applications will have to be changed considerably to be adapted to ATM and make subtle use of resources.
- ATM does not provide security. This will have to be handled in higher layers.
- As public ATM network deployment is still very slow, the connectivity in the public WAN area is bad today.
- ATM generates quite a lot of overhead (~ 20 %).

5. IP/ATM Co-Existence

Given the vast installed base of LANs today, the variety of LAN based applications and the network layer protocols operating on these networks, the key to the success of ATM in the short and medium term will be its ability to allow for interoperability between itself and these technologies. To enable the connectivity between ATM and existing LANs it is essential to use the same network layer protocols (such as IP, IPX) to provide a uniform network view to higher level protocols and applications.

Today, there are two standards available to run the predominant IP protocol over ATM: ATM Forum's LAN-Emulation (LANE) and IETF's "Classical IP over ATM". Both use the so-called *overlay model*, where IP addresses are mapped to ATM addresses. ATM is only used as a very fast packet transmission system and neither LANE nor Classical IP over ATM can therefore exploit ATM's QoS support as the IPv4 layer hides all the good features of ATM from higher layers. Moreover both technologies can establish ATM end-to-end VC connections only inside a 'subnet' (LIS or VLAN) and require IP routers for traffic across 'subnets', with the routers becoming potential performance bottlenecks. LANE and Classical IP over ATM are presented in Section 5.1.

There are also several ongoing activities in the ATM Forum and the IETF to enhance their overlay protocols to make better use of ATM. NHRP and MPOA are discussed in Section 5.2.

However there are solutions available already today, which can bring QoS to IP based applications by supporting end-to-end ATM VC connections on a per flow basis and across subnet borders. Section 5.3 introduces Arequipa, providing application requested end-to-end ATM connections with QoS for IP based applications, and some of the approaches that combine the label switching technology of ATM with network layer routing (IP Switching, Tag Switching) while avoiding the usage of ATM addressing, routing and signalling altogether.

5.1 Co-Existence without QoS Support

There are two fundamentally different ways of running network protocols over ATM networks. One method is the *native mode* operation, where network layer addresses are mapped directly to ATM addresses and network layer packets are sent directly across the ATM network, the other method is *LAN Emulation*.

5.1.1 LAN Emulation (LANE)

5.1.1.1 General overview

The ATM Forum specified LAN Emulation (LANE) in order to accelerate the deployment of ATM in the local area while native mode operation is still under definition. LANE offers a solution to the problem of running predominant local area protocols like Ethernet and Token Ring transparently over an ATM network. The current version of LANE can be found in [19].

LANE emulates a bridged LAN on top of an ATM network by offering a service interface to network layer which is *identical* to that of existing LANs (e.g. IEEE 802.3 Ethernet or 802.5 Token Ring) and it sends data across the ATM network using appropriate LAN MAC encapsulation. In brief, LANE makes an ATM network look and behave like an Ethernet or Token Ring, albeit a fast one. The big advantage of emulating a LAN is, that all network layer protocols and applications can be used without any modifications.

Today, LANE protocol software is widely available on ATM hosts (either implemented in the operating systems or on ATM network interface cards (NIC)) and on LAN Switching Equipment. ATM switches are transparent for the operation of the LANE protocol. They do not need to be modified for the use of LANE, although some of the LANE server components could be implemented on them.

The main issues on emulating a LAN technology like Ethernet on an ATM network is address resolution, broadcast and data encapsulation. Address resolution from MAC to ATM addresses is solved by using a special protocol called LE_ARP between hosts and a special LANE entity known as the LES (LAN Emulation Server). Broadcasting is emulated by sending packets to another LANE entity known as the BUS (Broadcast and Unknown Server) which distributes the packets to all hosts. The LAN packets (e.g. Ethernet frames) are encapsulated in AAL5.

5.1.1.2 Architecture

In the upper part of Figure 9, the architecture of a standard bridged LAN environment is shown. On a shared medium LAN (such as Ethernet) all packets sent by one station travel to all other stations on the medium. Bridges are intelligent repeaters (layer 2) which try to avoid unnecessary forwarding of packets. The functionality of a LAN segment can be emulated by an ATM network running LANE (lower part of Figure 9). This emulated LAN segment is called an Emulated LAN (ELAN). Together with the remaining old LAN infrastructure it forms a virtual LAN (VLAN).

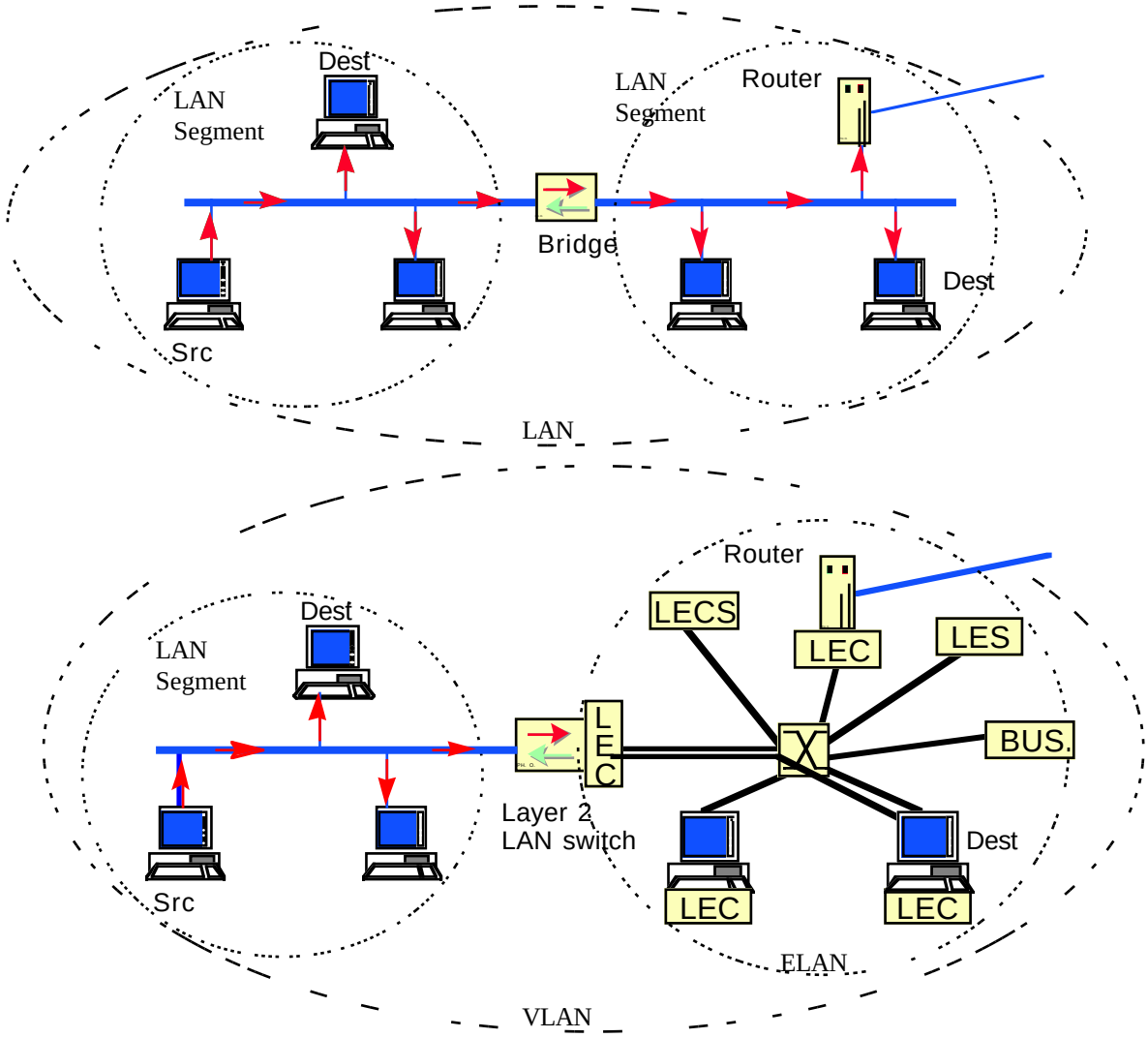

Figure 9: Classical LAN and Emulated LAN architecture

For the operation of LANE the following entities are needed:

- **LEC** (LAN Emulation Client)
  A LEC runs on every host. It provides a standard LAN interface to upper layers. The LEC issues address resolution requests and performs data encapsulation and forwarding.
- **LES** (LAN Emulation Server)
  There is a single LES per ELAN. It registers the mapping of MAC to ATM addresses and replies to or forwards address resolution requests.
- **BUS** (Broadcast and Unknown Server)
  There is a single BUS per ELAN. It emulates broadcasting by forwarding packets to all known ATM addresses on the ELAN.
- **LECS** (LAN Emulation Configuration Server)
  There is a single LECS per domain, used for the configuration of several ELANs.

Several Virtual Channel Connections are needed between these entities. Figure 10 shows the VCCs in a 2 host ELAN.

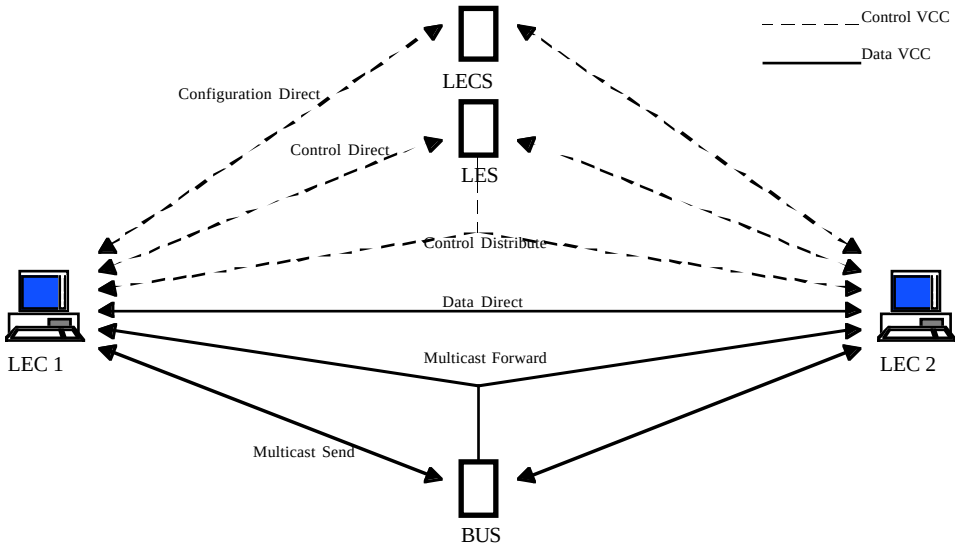

Figure 10: ATM connections for LANE

All these VCCs are established by signalling (SVCs). The VCCs are either UBR or ABR.

5.1.1.3 LANE procedures

LANE Configuration:

- LEC establishes the Configuration Direct VCC to LECS.
- LEC learns from LECS the ATM address of LES over the Configuration Direct VCC.
- LEC sets up the Control Direct VCC to LES and registers its ATM and MAC address in LES.
- LES adds LEC as a leaf to its point-to-multipoint Control Distribute VCC.
- LEC learns ATM address of BUS by using LE_ARP to LES for the MAC broadcast address.
- LEC sets up the Multicast Send VCC to BUS.
- BUS adds LEC as a leaf to its point-to-multipoint Multicast Forward VCC.

LANE Operation:

- LEC1 wants to send to LEC2, but only knows its MAC address.
- LEC1 uses LE_ARP request to LES to map LEC2's MAC address to its ATM address.
- While waiting for the reply, LEC1 sends packets to BUS, which floods it to all connected LECs.
- After receiving the LE_ARP response LEC1 sets up the Data Direct VCC to LEC2.
- Before sending on the Data Direct, LEC1 has to send a flush to BUS to make sure that all packets previously sent to LEC2 over BUS were delivered (to preserve frame ordering).

5.1.1.4 Assessment

LANE is a good solution to interconnect legacy LAN equipment in a private network, exploiting ATM's fast transmission speed with minimal changes to LAN equipment and no changes at all to higher layer protocols and applications. Its a working solution for today and allows for a smooth integration from LAN to ATM in a corporate network.

LANE even introduces enhanced configuration flexibility and improved management compared to standard LANs with its concept of *virtual* LANs.

However LANE can not be the ultimate solution for a modern integrated services network because of the following limitations:

- LANE totally hides the QoS support of ATM with its emulation of a connectionless shared media technology.
- LANE is unable to run protocols in native mode.

- LANE is limited to a logical subnet (VLAN).
- All inter-VLAN traffic has to pass through routers even if direct ATM connectivity would be possible. These routers are likely to become bottlenecks.
- LANE address translation is very inefficient because addresses are translated from Layer 3 addresses to MAC addresses to ATM addresses, using two different address resolution mechanisms.
- LANE operation needs a lot of connections, limiting the number of stations that can be attached to an emulated LAN
- LANE has not recovery mechanisms for the server, thus it does not foresee the possibility to define backup LES and BUS to manage the VLAN in emergency conditions.
- LANE has limit on the MTU size.

5.1.2 CLassical IP over ATM (CLIP)

5.1.2.1 General description

A solution to overlaying IP networks on ATM networks is the so-called *Classical IP over ATM*, specified by the IP-Over-ATM working group of the IETF and described in detail in RFC 1577 [20].

The *Classical IP* model refers to a network where hosts are organized in subnetworks sharing a common IP address prefix, where the ARP is used for IP address to MAC address resolution and where communication across subnetworks goes through routers. Preserving the classical IP model on ATM means that ATM is used as a direct replacement for the "wires" and local LAN segments connecting IP end-stations ("members") and routers operating in the "classical" LAN-based paradigm.

The Classical IP over ATM specification defines classical IP and ARP in an ATM network environment configured as a Logical IP Subnetwork (LIS) as illustrated in Figure 11. It does not describe the operation of ATM Networks in general.

It is the goal of RFC 1577 to allow compatible and interoperable implementations for transmitting IP datagrams and ATM Address Resolution Protocol (ATMARP) requests and replies over the ATM Adaptation Layer 5 (AAL5).

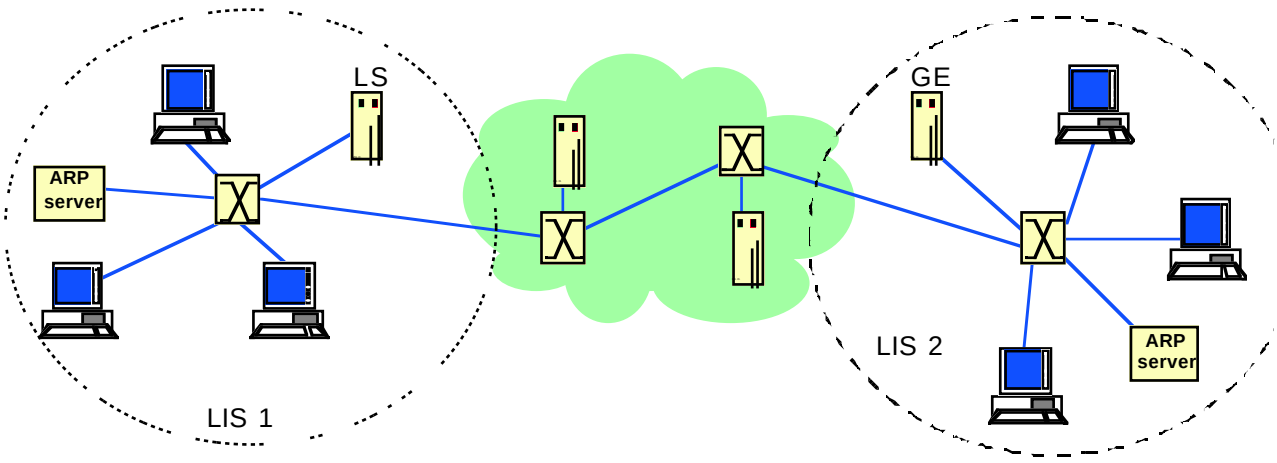

Figure 11: Classical IP over ATM network architecture

Data transmission in Classical IP over ATM is based on the virtual connection (VC) switched environment provided by ATM networks. The VCs can either be established by management (PVCs) or by signalling (SVCs). Each VC is directly connecting two IP members within the same LIS and carries all IP data flow between them.

The ATM connections could in principle be of any type (CBR, UBR, VBR, ABR), but only CBR and UBR is used in today's implementations..

5.1.2.2 Encapsulation of IP Datagrams

IP packets are transmitted using AAL5 with a maximum packet size (MTU) of 9180 bytes. Additionally, when using SVCs, IP packets must be encapsulated with LLC/SNAP [21] and the SETUP signalling messages to establish these SVC must carry Lower Layer Information (BLLI) indicating that the packets should be delivered to the LLC entity [22].

5.1.2.3 Address Resolution Mechanisms

When SVCs are used for transmission, special address resolution mechanisms are needed to map IP addresses to ATM addresses and vice versa. Similar to classical IP networks where ARP [4] and InARP [23] are used to map between IP and MAC addresses, Classical IP over ATM defines ATMARP and InATMARP services to map between IP and ATM addresses. For example, if Host A wishes to send IP datagrams to Host B it needs to have the ATM address of Host B to be able to establish a switched VC using signalling. For this IP to ATM address resolution, the ATMARP service is used. The originating host sends an ATMARP request to a special network entity, the dedicated ATMARP server of the LIS. The ATMARP server, knowing the IP and ATM addresses of all hosts and routers in its LIS (see below), maps the provided IP address of Host B to the corresponding ATM address and sends it back to Host A. Host A can then establish an SVC to Host B using normal signalling procedures. Host B then uses InATMARP procedures on this newly established connection to learn the IP address of Host A.

When hosts are connected by PVCs, they may use a preconfigured table to map IP addresses to VCs but they have a mechanism for resolving VCs to IP addresses via InATMARP for new VCs.

Each host must know its own IP and ATM address(es) and must respond to address resolution requests appropriately. It must also be configured with the ATM address of an ATMARP server (for SVCs only) located within the LIS (there is only one server per LIS). At power-up a host establishes a connection to the server. On each new incoming connection the ARMARP server send an InATMARP request and registers the reply. The reply contains the information for the ATMARP server to build its ATMARP table cache. This information is used to generate replies to the ATMARP requests it receives.

Because ATM does not support broadcast addressing, there is no mapping from IP broadcast addresses to ATM broadcast services. This is currently also true of multicast address services, although an Internet draft for multicast support already exists [24].

All hosts as well as the server must maintain an ATMARP table. A table entry contains the IP address, ATM address and VCI/VPI of a connection together with encapsulation information and a timestamp. Hosts must refresh the entries at least every 15 minutes and the server must refresh the entries at least every 20 min. Connections are released after a certain idle period.

It is important to stress the fact that the address resolution mechanisms of Classical IP over ATM can only be used inside a single LIS and not across LIS borders.

5.1.2.4 Routing

Classical IP over ATM uses exactly the same end-to-end routing architecture as the classical IP network. As the classical IP network uses ARP protocols and tables for the routing inside of the subnetwork, so does Classical IP over ATM use ATMARP protocols and tables for the routing inside of a LIS. For communication across LIS borders routers are needed in the same way as when crossing subnet borders in classical IP networks. This leads to the necessity of using several hops over IP routers across an ATM networks for traffic between hosts in different LIS (see Figure 12).

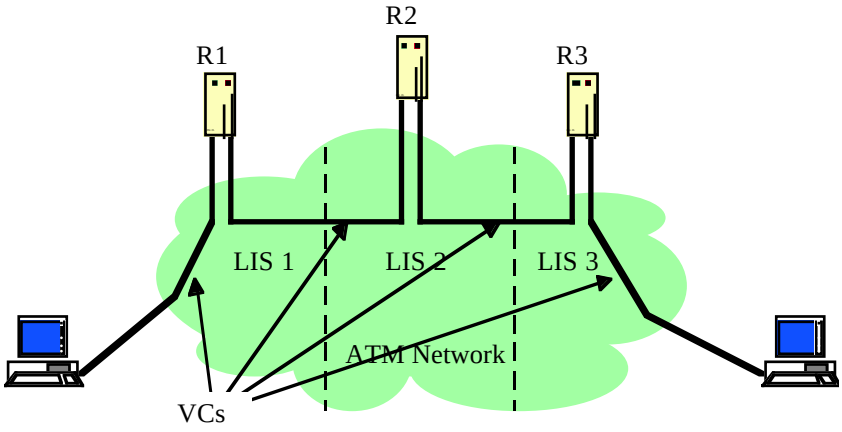

Figure 12: routing for traffic across LIS borders

5.1.2.5 Assessment

The main advantage of using Classical IP over ATM is its full compatibility with normal IP, enabling the vast set of higher layer protocols and applications to run transparently over ATM while making use of ATM's high bandwidth availability. Another advantage is that Classical IP over ATM allows easy integration of IP based services with other ATM services (e.g. voice).

The major shortcoming of Classical IP over ATM is that it can not benefit from ATM's inherent end-to-end QoS guarantees for the following reasons:

- Direct ATM connections can only be established inside a LIS but not across LIS borders.
  Because of using the classical IP routing mode (address resolution is limited to a LIS) IP traffic between hosts on differing LISs always flows via one or more intermediate IP routers who can only provide best effort delivery on IP level. This results in a concatenation of ATM connections even though it may be possible to open a direct ATM connection between the two hosts, thus preempting end-to-end QoS. In other words, IP packets across LIS borders hop several times through the ATM network instead of using one single hop.
- All IP data flowing between two hosts shares the bandwidth of a single VC.
  Having only one shared VC between two hosts makes it impossible for individual applications to get a QoS guarantee for their specific data flow.

Other shortcomings of Classical IP over ATM are that neither multicast nor anycast is supported, that IP layer implementations need to be adapted to interface with ATM directly and that it is necessary to deploy routers with ATM interfaces in every LIS. Furthermore there is no default path to forward IP datagrams before a connection is established, resulting in a high delay for the passing of the first datagram.

Unlike LANE Classical IP over ATM does not allow to use much of the old legacy LAN equipment, but it offers a more appropriate MTU size (larger).

5.2 QoS Support by Emerging Standards

Both the IETF and the ATM Forum are aware of the shortcoming of their respective solutions of running IP over ATM (CLIP, LANE) and try to solve them by defining new additional standards. NHRP, under definition in IETF, tackles the extra-hop problem (router hops are required for traffic across LIS instead of direct ATM connections) to provide end-to-end ATM connectivity and bring the QoS features closer to IP based applications while generalizing on Layer 2 (IP over any layer 2). MPOA, under definition in the ATM Forum, defines a way to emulate a routed protocol over ATM and also addresses the extra-hop problem but generalizes on Layer 3 (any layer 3 over ATM).

5.2.1 Next Hop Resolution Protocol (NHRP)

5.2.1.1 General description

The IETF is generalising its approach to support IP (and other internetworking protocols) not only over ATM but over all kinds of Non-Broadcast Multiple-Access (NBMA) networks, such as ATM, Frame Relay or X.25. For this purpose, the IP over NBMA (ION) working group was formed as a successor of the Routing on Large Clouds (ROLC) and the IP over ATM (ipatm) working groups. The Next Hop Resolution Protocol (NHRP) was defined by ION as a key element of supporting IP over NBMA. NHRP is currently only an Internet Draft [29].

NHRP addresses one of the key problems in NBMA networks, namely the problem of stations communicating over a Non-Broadcast Multiple-Access (NBMA) subnetwork, that are not on the same LIS. The NHRP protocol allows the internetworking layer addresses and NBMA addresses of suitable "NBMA next hops" toward a destination station to be determined.

As we already pointed out in the description of Classical IP over ATM (see section 5.1.2), the address solving problem arises when the stations are in the same NBMA network, but not in the same LIS. In fact, in this scenario, classical address resolution as described in RFC1577 [20] and RFC1209 [30] does not work, because it can only discover a router that is a member of multiple LISs, and packets can hop several times through the NBMA network instead of using one single hop. NHRP solves this problem with the definition of an inter-LIS address resolution mechanism, providing the source station with a "short-cut" routing, that allows to communicate through the NBMA network without having to involve intermediate routers.

In this sense NHRP is not a routing protocol, but just an inter-LIS address resolution mechanism that makes use of network layer routing in resolving the NBMA address of the destination. Therefore NHRP does not replace existing routing protocols, that are still used to determine the source path (other means than routing can be used to do it, for example, static configurations).

NHRP replaces the concept of LIS with the concept of Local Address Groups (LAGs). LAGs were introduced in [31] to extend IP architecture, that limits direct communication between hosts with the same subnet, to large data network. LAGs are identified by an IP address prefix, and group hosts and routers with different subnet. As described in [29], for NHRP the essential difference between using the LIS or the LAG models is that while with the LIS model the outcome of the “local/remote” forwarding decision is driven purely by addressing information, with the LAG model the outcome of this decision is decoupled from the addressing information and is coupled with the Quality of Service and/or traffic characteristics. This implies that two stations that are on the same NBMA, but that are not necessary on the same LIS, can directly communicate being part of the same LAG, as illustrated in Figure 13.

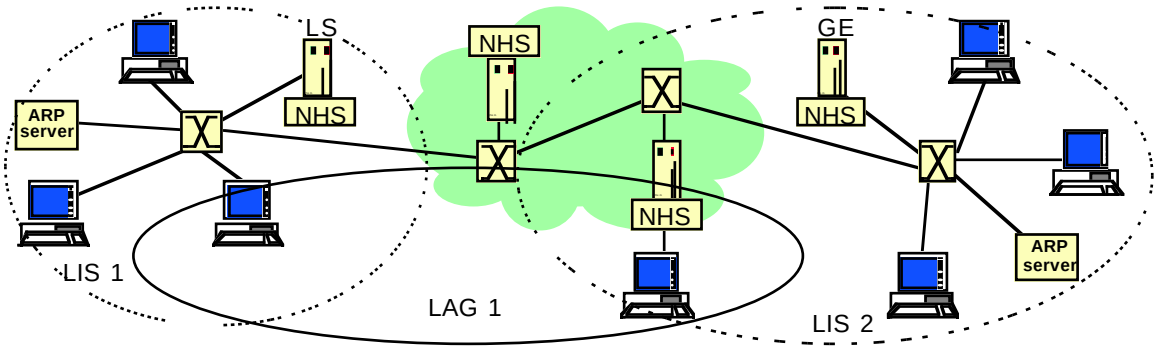

Figure 13: the Local Address Group (LAG) concept

5.2.1.2 Protocol overview

For NHRP operation there has to be one Next Hop Server (NHS) in every LIS. All hosts on a LIS register their NBMA and internetwork layer (e.g. IP) address with their NHS when booting.

Assume a Host S wants to send an internetwork layer packet (e.g. IP) to Host D which lies outside its LIS. To resolve the NBMA address of D, S sends a next hop *Resolution Request* to its NHS. The NHS checks whether Host D lies in the same LIS (is served by the same NHS). If the NHS does not serve Host D, the NHS forwards the request to the next NHS along the routed path. Using this algorithm the request is passed on from NHS to NHS and eventually arrives at the NHS that serves Host D. This NHS can resolve Host D's NBMA address and sends it back to Host S in a next hop *Resolution Reply* either along the routed path or directly. If it is sent back along the routed path, intermediate NHSs can optionally store the address mapping information for Host D contained in the Resolution Reply to answer subsequent Resolution Requests. Using this mechanism NHRP provides S with the NBMA address of D, if D is directly attached to the NBMA, or in the other case the address of an egress router at the edge of the NBMA which has connectivity to D. Host S and Host D may choose to cache the address mapping.

Host S can choose to either drop the packet triggering NHRP, retain it until the arrival of the Resolution Reply or forward the packet along the routed path towards Host D.

5.2.1.3 Use of NHRP

Issuing an NHRP request would be an application dependent action [31], in particular because NHRP allows the special features provided by the NBMA to be used. Thus, when a "cost" is associated with NBMA connections, there is an evident advantage in using NHRP short-cuts, i.e. only one connection across the NBMA. For example, when the NBMA network is ATM and the application requests QoS guarantees, the short-cut routing of NHRP helps to establish a direct VC in the ATM domain across several IP subnets, allowing the application to benefit from the QoS features of ATM.

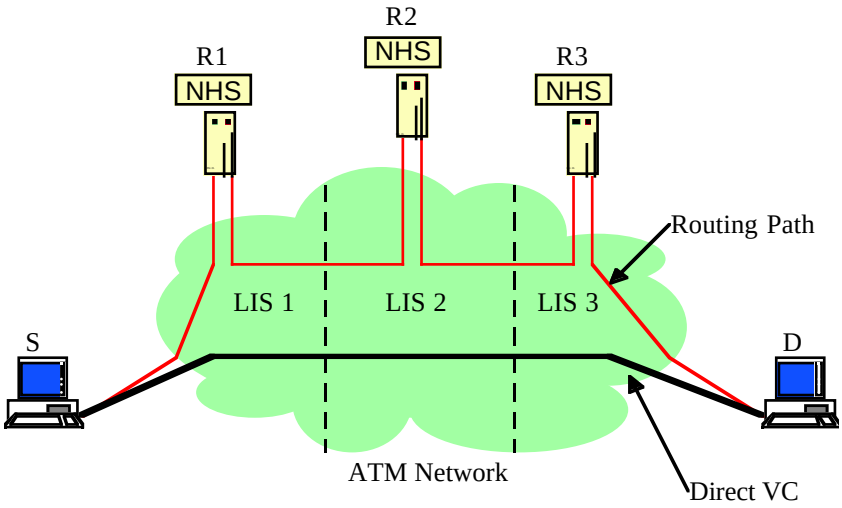

Figure 14: NHRP established direct ATM connection across LIS borders

For this reason, the Multiprotocol Over ATM (MPOA) Working Group of the ATM Forum has decided to use NHRP for resolving the ATM addresses of MPOA communications where the destination does not belong to the same Internetwork Address Sub-Group of the source [32], as illustrated in Figure 14.

5.2.1.4 Assessment

The main advantage of NHRP is that it can solve the multiple-hop problem through NBMA networks by offering inter-LIS address resolution, thus enabling the establishment of a single-hop connection through the NBMA network. If the NBMA network is ATM, this means that by using NHRP a single direct VC can be established across several LIS, bringing QoS to the IP data flow between the VCs endpoints. But NHRP can only achieve this if the routed path lies entirely within the NBMA network and only under the conditions that NHRP is supported on all routers along the routed path. Furthermore it is also important to note, that even if a direct connection can be established through the NBMA network, it will be shared by all IP traffic between the two endpoints, which means that it does not bring QoS to an individual application.

Another problem with NHRP is that stable routing loops may occur, if NHRP initiating and responding stations are routers, which are additionally connected over another network. Avoiding these routing loops imposes restrictions on the network configuration. But there is already work in progress [33] to augment NHRP to solve this problem.

Another negative effect that could arise with NHRP *Resolution Request* is the domino effect. This occurs when a router originates a NHRP *Resolution Request* for a transit packet (a packet arriving over one of its NBMA attached interfaces). If the router forwards this data packet without waiting for an NHRP transit path to be established, then the next transit router receiving the packet can originate its own NHRP *Resolution Request* and forward the packet, and so on. One solution proposed to solve this problem is that a router does not generate NHRP *Resolution Request* for transit packets, but only for packets on its non NBMA interfaces.

Deployment of seamless NHRP functionality requires additional software on all hosts and routers connected to the NBMA network.

The current NHRP specification works only for unicast communication, it does not suit a broadcast or multicast setting.

NHRP is only a draft and is far from being generally deployed.

5.2.2 Multiprotocol over ATM (MPOA)

5.2.2.1 Motivation

The ATM Forum's Multiprotocol Over ATM (MPOA) subworking group is defining an approach to support seamless transport of layer 3 protocols across ATM networks. Multiple layer 3 protocols are to be supported, such as IP, IPX, Appletalk, etc.

MPOA is extending the VLAN beyond what was defined in LANE based VLANs, addressing the well known shortcomings of LANE that router hops are required for VLAN interconnection and its inability to run protocols in native mode, which could exploit ATM's QoS features. In other words, MPOA tries to offer transparent emulation of routed protocols over ATM network, much the same as LANE offers transparent emulation of a LAN protocol over ATM network. MPOA provides end-to-end Layer 3 connectivity between hosts attached to the ATM fabric and hosts attached to legacy subnetwork technology. MPOA operates at layer 2 and 3, but uses LANE for layer 2 forwarding.

MPOA was built with the following design goals in mind:
- Allow MPOA devices to Establish Direct ATM connections
- No significant changes to installed Bridges, Routers and Hubs
- Integrate with LAN emulation
- Support Network Layer Multicast and Broadcast
- Support Auto Configuration at ATM hosts
- Separate Switching from Routing

Much as the IP oriented IETF is trying to run only IP over all underlying technologies (ATM being only one of them), the ATM Forum tries to run all kind of Layer 3 (IP only one of them) protocols over only ATM. Where "IP over ATM" is concerned the two standardisation bodies converge and the IP version of MPOA can be considered the unification of Classical IP over ATM (together with MARS and NHRP extensions) and LANE.

So far the ATM Forum produced a MPOA Baseline document [32].

5.2.2.2 The MPOA reference model

The basic unit of organisation within MPOA is the Internetwork Address Sub-Group (IASG). It is defined as a range of internetwork layer addresses summarized into an internetwork layer routing protocol. In the case of IP this is essentially a *subnet*.

An IASG will contain a number of devices acting as MPOA servers and clients as described in the MPOA reference model (Figure 15). Servers are those devices providing layer 3 co-ordination, address resolution, route distribution and broadcast/multicast forwarding. Clients are users of the MPOA services.

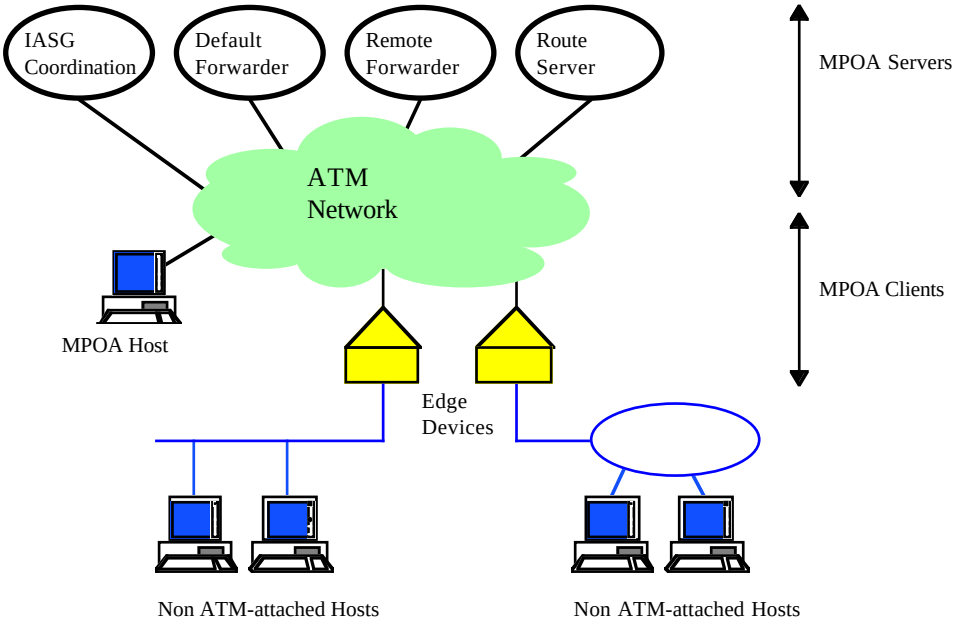

Figure 15: MPOA reference model

MPOA Clients are:

- **MPOA Hosts**: hosts that are directly attached to ATM, running MPOA protocol stack
- **Edge Devices**: physical devices that are capable of forwarding packets between legacy LAN interfaces and ATM interfaces at both Layer 2 and Layer 3. However they do not run layer 3 routing protocols to get the information for the Layer 3 packet forwarding, but they query the Route Server for this information.

The services offered by MPOA Servers can be classified in the following functional groups:

- **ICFG (IASG Coordination Functional Group):** coordinates the distribution of an IASG across multiple traditional LAN ports and/or ATM connected hosts, it is responsible for the configuration of the IASG
- **RSFG (Route Server Functional Group)**: runs layer 3 routing protocols, provides address resolution and route distribution
- **DFFG (Default Forwarder Functional Group):** forwards traffic within an IASG if no direct client to client connection exists and performs the Multicast Server Function (MSF) within the IASG
- **RFFG (Remote Forwarder Functional Group):** forwards traffic between IASGs

5.2.2.3 MPOA architecture

Typically the MPOA server functionality is split among two physical entities, the Route Server and the IASG Coordinator. The IASG Coordinator provides ICFG and DFFG functionality. The Route Server provides RSFG and RFFG functionality.

MPOA Hosts have direct VCs to the IASG Coordinator and to the Route Server. Edge Devices, Bridges and LANE Hosts connect to the server entities over LANE, implying that the MPOA servers and these devices all run a LEC.

5.2.2.4 MPOA procedures

Procedures in MPOA are highly complex. Nevertheless a simplified description is given which relates to Figure 16.

When initializing, all MPOA Hosts and Edge Devices announce their own Layer 3 and ATM addresses and the layer 3 addresses reachable through them to the IASG Coordinator and the Route Server. In parallel normal LANE initialisation takes place.

When a MPOA host desires to know how to contact another host over ATM it issues an address resolution query to ICFG. If the destination host is a MPOA host within the same IASG, ICFG can reply with its ATM address. If the destination host is in another IASG, the request will be passed among the RSFGs across IASG borders. In the destination host's IASG a RSFG/ICFG knows the ATM address of the destination host and can reply to the address resolution query. In either cases the source host can then establish a direct ATM connection to the destination host. Note that this functionality is identical to NHRP and indeed MPOA relies on this protocol. But in addition to the functionality of establishing a direct ATM connection, MPOA offers the passing of packets before the ATM connection is established by sending it from the source host over DFFG and several RFFG to the destination host along the routed path.

Now consider an Edge Device trying to send packets to another host over ATM. The edge device first looks at the MAC destination address. If it is not the MAC address of a router within the IASG, it has to remain inside the IASG and the Edge Device uses LANE either to send it directly if it knows the MAC to ATM address mapping (e.g. destination is a LANE host) or to send it to ICFG for forwarding. If the MAC address is the MAC address of a router, the Edge Device looks at the internetwork address contained in the packet. If it knows the internetwork to ATM address mapping (e.g. destination is a MPOA host) the Edge Device can forward it directly. If the internetwork address is unknown, the Edge Device asks the Route Server for an internetwork to ATM address resolution. In the latter case the Edge Device has the same behaviour like the MPOA host described in the previous paragraph.

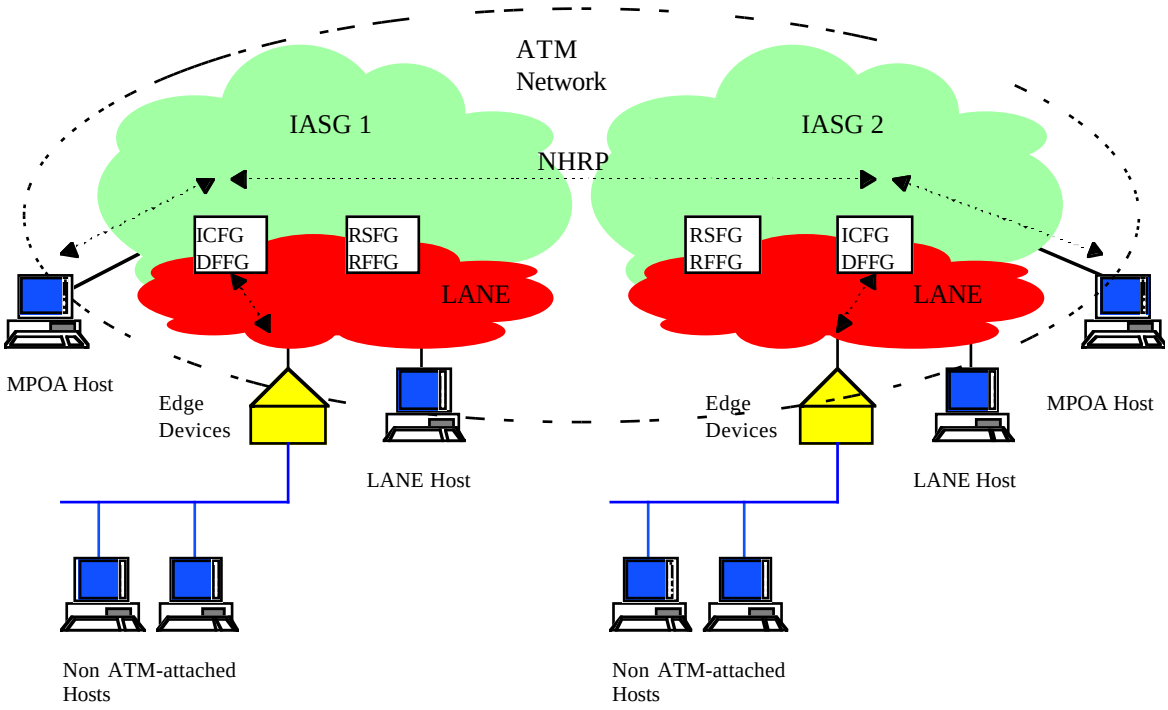

Figure 16: MPOA architecture

5.2.2.5 Assessment

MPOA is a very complex technology and the work in the ATM Forum has only started and is far from being completed. IP might be worth the complexity because it is so widely used, but it can be doubted if this holds true for other layer 3 protocols as well .

Nevertheless, the MPOA model is a very promising technology providing the following benefits:

- MPOA provides the connectivity of a fully routed environment, supporting even multicast and broadcast at layer 3.
- MPOA takes maximum advantage of ATM:
  - because it offers direct ATM connection between MPOA devices, without intermediate hops
  - because it supports Native ATM, exposing QoS to layer 3 protocol stacks
- MPOA reduces infrastructure costs by defining a new network architecture. Instead of deploying common routers with both the functionality of switching, which is very cheap as it can be done in hardware, and route computation, which is rather expensive as it needs to run on a high performance platform, the switching is distributed in Edge Devices and there is only a single, centralised router
- MPOA provides an universal approach for layer 3 protocols over ATM
- MPOA easily integrates with LANE

Apart from its complexity, a disadvantage of MPOA is that host protocol stacks have to be changed.

5.3 QoS Support with Existing Technologies

This section discusses some of the solutions available today to bring ATM's high speed and QoS support to IP based applications. Most of these solution were born as proprietary solutions of router vendors or educational institutions and then put forward to the IETF to make them standards (RFC).

Section 5.3.1 discusses Arequipa, an extension to Classical IP over ATM, which allows applications to request their own SVC with guaranteed QoS by bypassing the IP layer during connection establishment.

The rest of Section 5.3 discusses two of the various solutions of how to use the fast Layer 2 label switching of ATM in conjunction with network layer routing. The basic idea behind all of these technologies is to increase the packet forwarding performance of routers by replacing slow and expensive network layer forwarding decisions with fast, low cost Layer2 label-swapping based forwarding (cut-through packet forwarding) while at the same improving routing functionality, scalability and flexibility. If these technologies are seamlessly deployed in an ATM based network, end-to-end ATM VC connection with guaranteed QoS can be established for IP traffic, without having to use ATM addressing, routing and signalling like in the overlay model. The IETF Working Group MPLS (Multiprotocol Label Switching) is currently working on unifying and generalizing the different approaches which vary in such things as the type of used labels, the trigger event for label binding, the way how labels are distributed in the networks and the protocols they support. Section 5.3.2 introduces the concept of IP Switching because this was the first proposal in this area and Section 5.3.3 presents Tag Switching, which is probably the most advanced solution in this area today. Other related approaches like Cell Switch Router (CSR, Toshiba), Aggregate Route-Based IP Switch (ARIS, IBM) or Switching IP Through ATM (SITA, Telecom Finland) are not discussed in this paper.

5.3.1 Arequipa extension to Classical IP over ATM

5.3.1.1 General description

The Arequipa (Application REQuested IP over ATM) protocol is a mechanism which allows IP based applications to request their own SVCs with guaranteed QoS. It was developed by EPFL (a member of the ACTS-EXPERT project) as an extension to CLIP, and is described in RFC2170 [25].

Arequipa is a mechanism which allows applications to establish end-to-end ATM connections under their own control, and to use these connections at the lower protocol layer to carry the IP traffic of specific sockets, as illustrated in Figure 17.

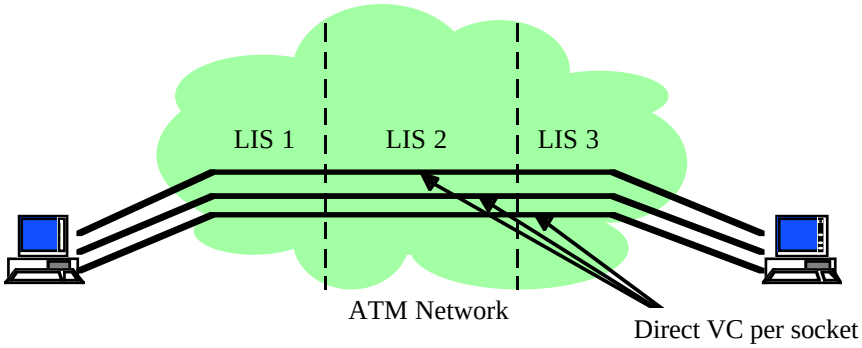

Figure 17: Arequipa established VCs across LIS borders

Unlike the connections set up by Classical IP over ATM or by LANE, which are shared by the entire IP traffic flow between the connection endpoints, Arequipa connections are used *exclusively by the applications that requested them*. The applications can therefore exactly determine what QoS will be available to them.

Figure 17 illustrates that Arequipa connections are end-to-end, despite the LISs topology, in line with the extensions to IP architecture described in [31]. It shows also that each flow has its own connection with QoS requirements.

In its broadest sense, Arequipa offers the means to use properties of a network technology that is used to transport another network technology (e.g. IP on ATM) without requiring the explicit design and deployment of sophisticated interworking mechanisms and protocols.

Traditional protocol layering typically only allows access to functionality of lower layers if upper layers provide their own means to express that functionality. This approach can introduce significant complexity if the semantics of the respective mechanism are dissimilar. Also, if the upper layer fails to provide that interface, no direct access is possible and the lower layer functionality may be wasted or used in an inefficient way (e.g. if using heuristics to decide on the use of extra features). This is apparent in the case with the QoS functionality of ATM which is hidden by the IP layer when IP is run on top of ATM. Arequipa enables applications to exploit the hidden properties of lower layers by allowing applications to control them directly.

It is important to note that Arequipa coexists with “normal” use of the networking stacks, i.e. applications not requiring Arequipa do not need to be modified and they will continue to use whatever other mechanisms are provided. Moreover, although traffic between applications using Arequipa does not pass the normal routed IP path anymore, general IP connectivity may still be necessary, e.g. for ICMP messages or for traffic of other applications.

5.3.1.2 Protocol overview

Arequipa provides two new socket primitives to applications:

- Arequipa_preset(): opens an end-to-end SVC and sends all data from the socket over that connection
- Arequipa_expect(): allows incoming Arequipa connections in the reverse direction

Typically the server side of the application opens and binds a socket and then calls Arequipa_expect(), preparing the socket for incoming Arequipa connections. The client side opens a socket, calls Arequipa_preset() with the desired QoS and the server’s ATM address and port number and then connects the socket.

Note, that in order to establish the direct VC connection, the ATM address and port number of the server has to be known.

In the protocol stack Arequipa can be seen as a device. Figure 18 shows the protocol stack for Arequipa and the interaction for a Arequipa_preset call.

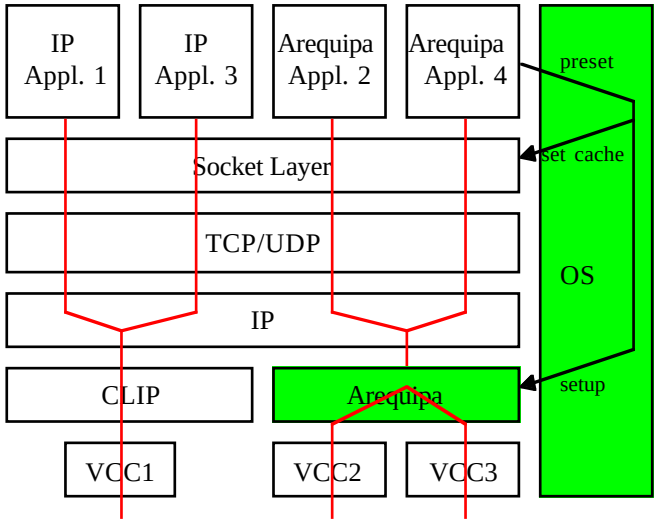

Figure 18: Arequipa in the protocol stack

5.3.1.3 Applicability

Arequipa is applicable, for IP and ATM, if the following two conditions are met:

- applications can control "native" connections over the lower layer communication media, that is that there has to be a signaling API which can be used by an application
- both IP and ATM allow communication between the same endpoints (or they share at least a useful common subset of reachable endpoints)

The next two conditions do not have to be met, but without them the use of Arequipa may be questionable:

- all IP traffic between a pair of hosts typically shares the same ATM SVC
- multiple lower layer connections are possible between a pair of endpoints

In order to simplify interaction with the protocol stack, Arequipa assumes that data sent to destinations for which no Arequipa lower layer connection has been established will be delivered by some default mechanism.

Note that despite its name (Application REQuested IP over ATM), Arequipa is not only limited to IP and ATM. The upper layer is typically IP or some similar protocol (e.g. IPX). The lower layer can be ATM, Frame Relay, N-ISDN, etc.

5.3.1.4 Application changes

TCP/IP based applications have to be slightly changed in their socket opening behaviour to enable them to run over Arequipa. Basically all that has to be changed is the calling of new socket functions Arequipa_preset() and Arequipa_expect().

There is already an Arequipa based application publicly available to demonstrate the power of the Arequipa approach. This is a Web-over-ATM application written by EPFL, which allows HTML pages to be downloaded with QoS guarantees.

5.3.1.5 Assessment

Arequipa has the following advantages:

- Arequipa enhances CLIP to allow IP based applications to make full use of ATM's QoS guarantees by allowing them to set up and control their own VC connections.
- Arequipa is a rather light software that only needs to be run on hosts and needs no network support like NHRP or RSVP.
- By establishing direct end-to-end connections routing overhead can be avoided.
- Arequipa is a solution that works and is available *today*.
- Arequipa is co-existent with the normal CLIP stack allowing "normal" and "Arequipa enhanced" applications to run simultaneously.

The only disadvantage in using Arequipa can be seen to be the fact that existing IP applications need to be "Arequipa enhanced" to be able to take full advantage of its features, though software changes are only minimal.

5.3.2 IP Switching

5.3.2.1 General Overview

An IP Switch is a hybrid between an ATM Switch and a gigabit router. Datagram forwarding is handled by an ATM switching fabric (as opposed to a router backplane) and routing is performed by traditional router software on an IP switch controller (Figure 20).

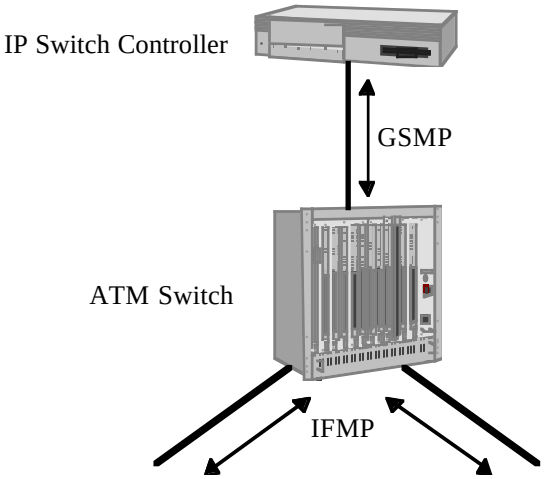

Figure 19: IP Switching Architecture

By using the high performance, low cost switching hardware of ATM together with the simple, well tuned IP software for addressing and routing, IP Switching combines the strength of both technologies (Figure 20).

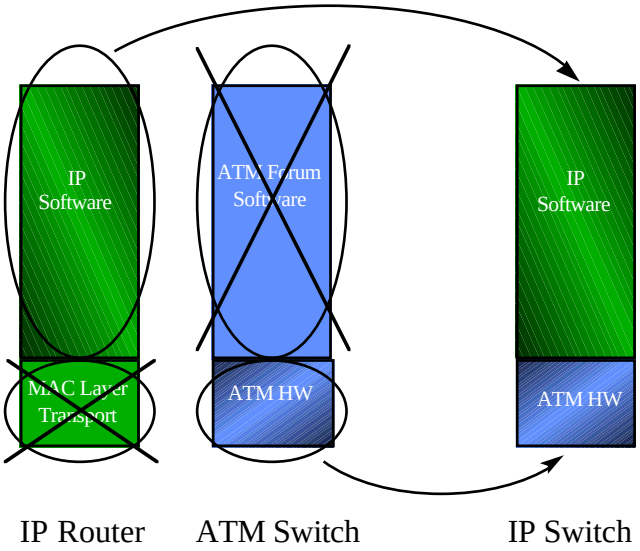

Figure 20: IP Switching concept

IP Switching uses flow classification to optimise the load on the IP switch controller. A *flow* is an extended IP conversation. More specifically, a flow is a sequence of IP packets sent from a particular source to a particular destination sharing the same protocol type (such as UDP or TCP), type of service, and other characteristics, as determined by information in the packet header. The switch controller identifies longer duration flows, as these

can be optimised by cut-through switching in the ATM hardware. The rest of the traffic continues to receive the default treatment - hop-by-hop store-and-forward routing.

5.3.2.2 Flow Classification

The main task of the flow classification process is to select those flows that are to be switched in the ATM switch, and those that should be forwarded packet by packet by the IP switch controller. The decision to switch flows directly through the ATM switch is called short-cut routing. Long duration flows are well adapted for such a short-cut routing. Short duration flows should be handled directly by the forwarding engine of the IP switch controller. Application information provides an approximate indication for flow duration. Multimedia traffic (voice, image, video-conferencing) is an example of long duration flows, whereas name server queries, are typically of short duration.

For the flows selected for short-cut routing, a VC must be established across the ATM switch and the association of flow and VCI label has to be communicated to the upstream IP switch in order that this switch can use a short-cut route. The Ipsilon Flow Management Protocol is a means to communicate this information, another solution would be to use RSVP.

5.3.2.3 Ipsilon Flow Management Protocol (IFMP)

IFMP [26][27] enables communications between multiple IP Switches or between hosts and IP Switches. It associates IP flows with ATM virtual channels and defines the format for flow-redirect messages and acknowledgements. IFMP is implemented in end stations, such as routers, shared-media hubs, LAN switches, or TCP/IP hosts equipped with an ATM NIC to connect directly to an IP Switch. On ATM links it uses a default VC (VPI 0, VCI 15). The ATM VCI for a specific IP flow is selected by the receiving end of the link. All packets of flows that have not been switched are forwarded hop-by-hop between IP switch controllers using the default VC.

At system start-up, each IP node sets up a virtual channel on each of its ATM physical links to be used as the default forwarding channel. IP data traffic from existing network devices flows into an upstream host, edge router, or IP Switch gateway equipped with an ATM network interface card (NIC) and IP Switching software.

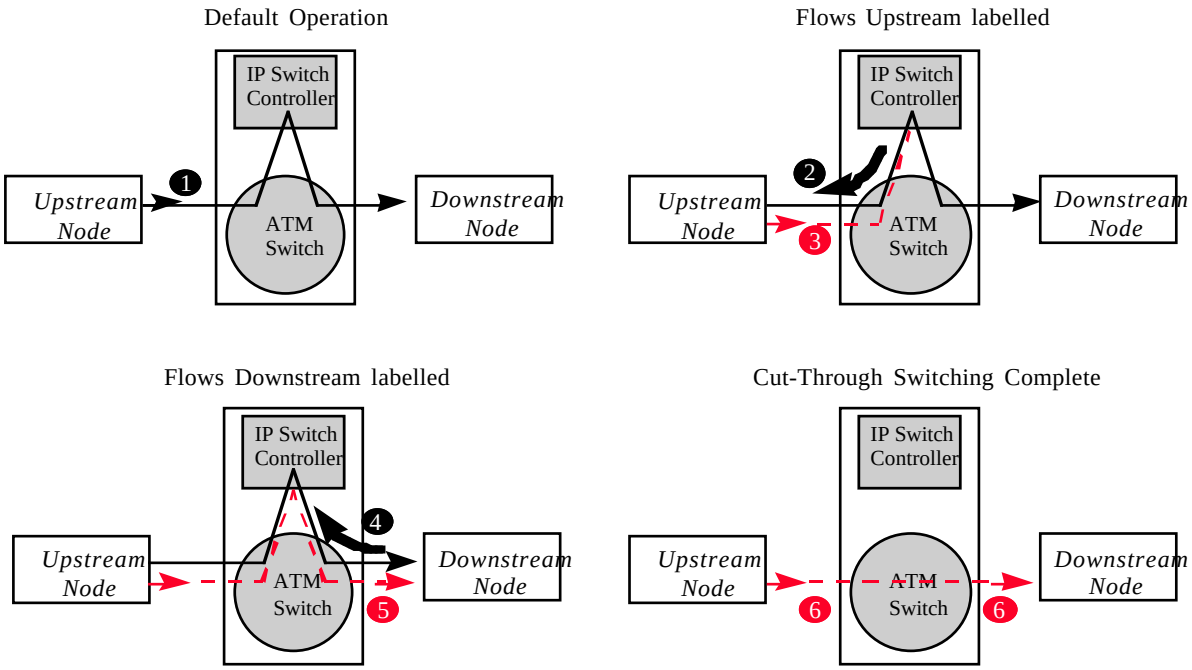

Figure 21: From default routing to cut-through ATM connection

An ATM input port inside the IP Switch receives incoming traffic from the upstream device on the default channel and sends it to the intelligent routing software of the IP Switch Controller (1). The ATM switch

hardware functions simply as a high speed I/O extension of the routing software. The IP Switch Controller forwards the packet in the normal manner over the default forwarding channel. It also performs flow classification, a decision-making process that enables IP Switches to optimise data traffic. Once a flow is identified, the switch controller asks the upstream node via IFMP to label that traffic using a new virtual channel (2). If the upstream node concurs, it selects a new virtual channel and the traffic starts to flow on this virtual channel (3). Independently, the downstream node can also ask the IP Switch Controller to set up an outgoing virtual channel for the flow (4). When the flow is isolated to a particular input channel and a particular output channel (5), the IP Switch Controller instructs the switch to make the appropriate port mapping in hardware, bypassing the routing software and its associated processing overhead (6). This design allows IP Switches to forward packets at rates limited only by the aggregate throughput of the underlying switch engine. First-generation IP Switches support up to 5.3 million PPS throughput. Further, because there is no need to reassemble ATM cells into IP packets at intermediate IP Switches, throughput remains optimised throughout the IP network.

5.3.2.4 General Switch Management Protocol (GSMP)

The control protocol used between the IP switch controller and the ATM switch is the General Switch Management Protocol (GSMP) [28]. This allows IP switching to be used with ATM switches from different suppliers. Different ATM switches are designed with different size, cost and functionality trade-offs, so a choice has to be made. GSMP can also support a standard ATM Forum control protocol stack instead of the IP switch controller software. Thus, a choice of network control software is possible for the same hardware.

GSMP is a simple master-slave, request-response protocol, and the switch issues a positive or a negative response, when the operation is complete. Unreliable transport is assumed between controller and switch for speed and simplicity. All GSMP messages are acknowledged, and the implementation handles its own retransmission.

GSMP runs on the default VC (VPI 0, VCI 15) over AAL 5 with LLC/SNAP encapsulation. The most frequent messages (connection management) are designed to fit into single cell AAL 5 packets. An adjacency protocol is used to synchronise states across the control link and to discover the identity of the entity at the far end of the link. There are five types of message: configuration, connection management, port management, statistics, and events.

GSMP has been implemented on at least eight different ATM switches. The code for the GSMP slave is about 2000 lines. A reference implementation is available. The measured performance of the GSMP slave on Ipsilon's IP switch is just under 1000 connection setups per second. This could be improved by hardware SAR support.

5.3.2.5 Assessment

IP Switching is describing an optimized and scaleable way of supporting IP over ATM. It makes use of the strength of both ATM and IP to increase the throughput of the Internet: ATM hardware offers fast speeds at relatively low prices; IP routing is much simpler than the complicated ATM addressing, routing and signalling protocols defined by the ATM Forum (UNI, P-NNI). Persistent flow traffic (e.g. file transfer) is typically worth the connection establishment delay and ATM overhead because once the direct VC is established only fast cell switching is performed by the network node without having to reassemble and analyze IP datagrams for routing. On the other hand, the delay and overhead of establishing an ATM connection does not make sense for short duration, non-persistent data flows (e.g. DNS lookup), which consists only of a few datagrams, where normal IP datagram routing is much better suited.

End-to-end QoS can in principle be achieved in a homogeneous, IP Switching equipped network. However QoS is only expressed with a priority for a flow and not with the usual ATM parameters for QoS. Furthermore it is important to note here, that it is not the application itself but the network that initiates the connection setup. This means that an application has no means to request a special QoS.

5.3.3 Tag Switching

5.3.3.1 General Overview

Tag Switching is a proprietary proposal by CISCO [34], [35]. Its objective is to increase router performance in WANs (for example in the global Internet or in the backbone of ISPs) by reducing the complexity of packet forwarding while providing better scalability and richer functionality to network layer routing. Unicast packet forwarding in an IP router involves searching in a table of IP address prefixes (called network layer reachability entries) for the prefix which has the longest match. Tag switching aims at replacing this operation as much as possible by a simple fixed length label lookup in hardware, exactly as is done with ATM or Frame Relay. This improves packet forwarding performance and introduces new functionality, increased scalability and more flexibility in the network layer routing.

Tag Switching consists of two components, the forwarding component, that uses the tag information in the packets and the Tag Information Base in the switch to perform fast packet forwarding, and the control component which is responsible for tag creation and distribution.

Tag Switching is not restricted to use IP as network layer protocol and ATM on Layer 2 but is a general approach applicable to any network layer and Layer 2 protocol.

5.3.3.2 Tags

Tags are short, fixed length labels, enabling Tag Switches to do simple and fast table lookups in hardware. Tag Switching does not define its own packet format it only adds a tag to an existing packet format. The tag information can be carried in a packet in a variety of ways. For example a 32-bit tag is added in front of a network layer package, which could be IPv4, IPv6, Appletalk or another format. Figure 22 shows this tag format. A tagged packet is carried on any layer 2 mechanism (e.g.: Ethernet, ATM) and is identified by a layer 2 protocol type (i.e., there would be an Ethertype defining unicast tagged packets, and another Ethertype for multicast packets). A minor difference compared to the 'tags' used in ATM and Frame Relay is the presence of a time-to-live field, which allows to use normal IP routing for tag distribution.

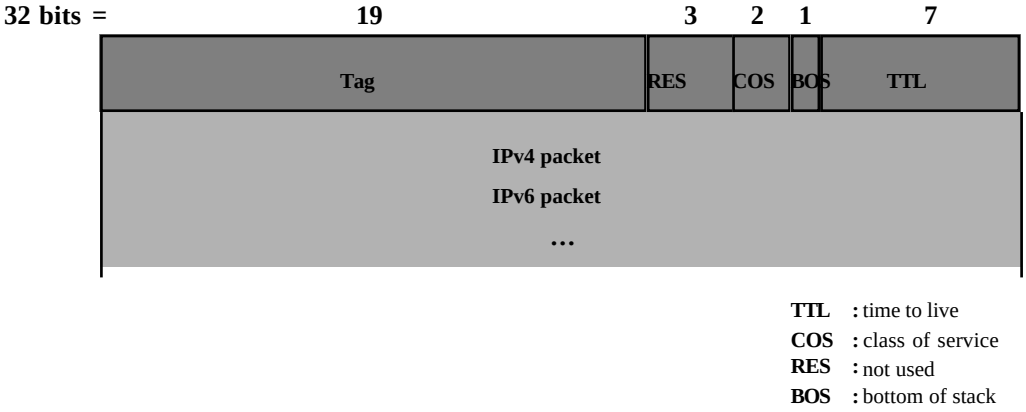

Figure 22: A tag format

Given the variety of ways to carry tag information enables the use of Tag Switching over any kind of media.

Tags may optionally be stacked. This enables aggregation of traffic flows. It can be used to speed up packet processing in backbones, and also to scale reservation mechanisms. Figure 23 shows a possible use of stacked tags in the Internet. Tag switch R1 adds an IGP tag to a BGP tagged packet to route it inside the domain. Tag switch R2 makes its forwarding decision solely on the IGP tag. Tag switch R3, the egress router of the domain, removes the IGP tag.

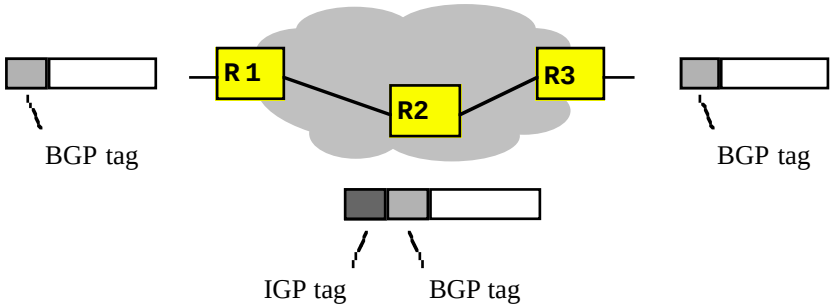

Figure 23: Stacked tags

5.3.3.3 Forwarding Component

The forwarding component of a tag switch is based on the notion of label swapping. Every tag switching node maintains a Tag Information Base (TIB), which is similar to ATM label swapping tables. If an incoming packet is tagged, then the Tag Information Base is searched for an exact matching entry. If one is found, then the Tag Information Base entry indicates the outgoing interface to which the packet should be forwarded, and the value of the new tag to be used. Unlike with ATM switching, if no entry is found, then the network layer information contained in the packet is used.

It is important to note that the forwarding component of Tag Switching is network layer independent.

5.3.3.4 Control Component

The control component of a tag switch is responsible for creating and distributing the tag binding information among tag switches.

In contrast to IP Switching where tag bindings are triggered by the detection of a persistent data flow (data traffic driven) Tag Switching uses topology driven tag binding, which means that a tag switch is populating its TIB with incoming and outgoing tags for all routes to which it has reachability.

Tag Switching supports a wide range of forwarding granularities to supports a wide range of forwarding granularities to provide good scaling characteristics and accommodate diverse routing functionality: at one extreme a tag could be bound to a group of routes, at the other extreme a tag could be bound to an individual information flow.

There are three permitted methods for tag allocation and TIB management:

- downstream tag allocation
- downstream tag allocation on demand
- upstream tag allocation

In downstream allocation a switch is responsible for creating tag bindings that apply to incoming data packets and receives tag bindings for outgoing packets from its neighbors (see Figure 24 (top)). Upstream allocation is the other way round.

There are two families of methods for tag distribution, namely tag distribution by explicit reservations and tag distribution based on destinations.

In tag distribution by explicit reservation, tags are distributed along with the reservation mechanism; if RSVP is used, then the value of the tag is part of the RESV message. This is very similar to the connection setup mechanism of ATM.

In tag distribution based on destination, the tags are distributed by the routing protocol. For this purpose, the tag switches also have to be routers for the protocols they support (IPv4, IPv6, Appletalk, etc.). Routing protocols are used to write the prefix entries, which are then associated with tags. Routing updates may piggyback the tags (distance or path vector protocols), or the tags may be distributed by a separate protocol called Tag Distribution Protocol [36] (link state protocols). Binding tag distribution together with routing is

much simpler than using the overlay model (like in IP over ATM). The presence of the TTL field in the tag avoids problems of temporary loops.

Figure 24 shows an example of tag distribution with a distance vector protocol and IPv4 address formats. It also shows the resulting Tag information Bases and the forwarding of tagged packets.

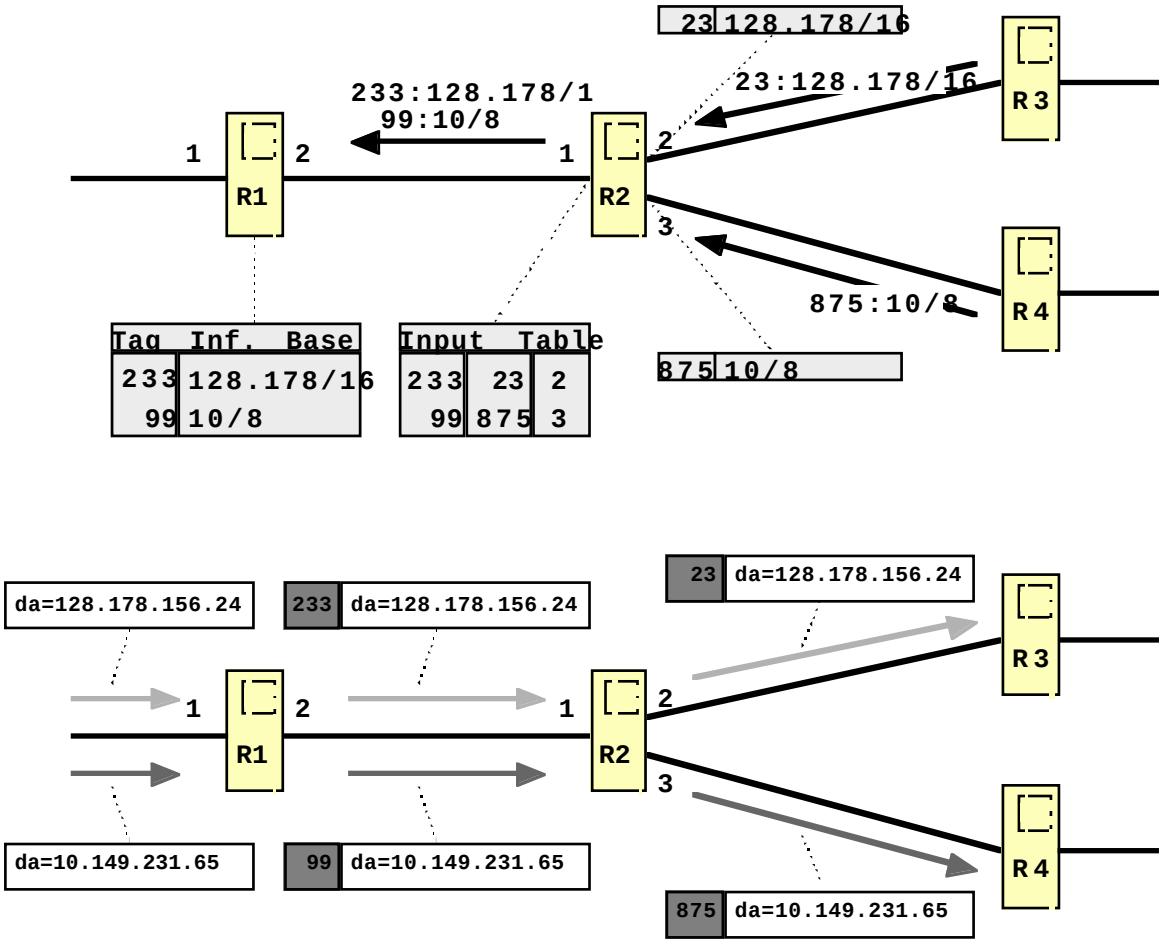

Figure 24: Tag distribution by routing updates (top) and forwarding of tagged packets (bottom)

5.3.3.5 Tag Switching with ATM

The characteristics of ATM switches require some specialized procedures and conventions to support tag switching (see [37]).

- Tags can be carried in the VCI field of ATM cells, or if two levels of tagging are needed in the VCI and VPI fields.
- The *downstream on demand* tag allocation procedure is used.
- ATM switches need to implement the control component of Tag Switching, have to actively participate as a peer in the network layer routing protocol and may need to support network layer forwarding.
- ATM tag switches are only allowed to be interconnected over conventional ATM switches if VP connections are used (only one level of tagging).
- To avoid cell interleaving an ATM tag switch needs to have several tags allocated with one route .
- The existence of the tag switching control component on an ATM switch does not preclude the ability to support the ATM control component defined by the ITU and ATM Forum on the same switch and the same interfaces. The two control components, tag switching and the ITU/ATM Forum defined, would operate independently.

5.3.3.6 Assessment

Tag Switching is a very powerful way of integrating the fast forwarding of cell-switching technologies with the simple addressing and routing of frame-switching technologies. The simplicity of the Tag Switching forwarding paradigm improves forwarding performance, while maintaining competitive price/performance. By associating a wide range of forwarding granularities with a tag , a wide variety of routing functions (destination based routing, multicast routing, QoS-based routing, hierarchy of routing knowledge) can be supported.

Tag Switching differs from IP Switching in that tags are never allocated based on flow analysis but based on the network topology. Because the network topology is quite static, topology-based tag allocation has a performance advantage over flow-based allocation. Another difference to IP Switching is that Tag Switching is a multiprotocol technology, which is neither bound to a particular network layer nor to a particular data link layer

If Tag Switching is used with IP and ATM, the whole huge ATM control plane as defined by the ATM Forum or ITU (UNI, P-NNI, etc.) can be replaced by the much simpler control component of IP Switching. A drawback of using Tag Switching with ATM is that the ATM tag switches have to participate as a peer in the network layer routing protocol and may even need to support network layer forwarding. If ATM Tag Switching is used in conjunction with a reservation protocol like RSVP it is possible to provide VC connections with guaranteed end-to-end QoS for IP flows or even applications in a homogeneous network.

Tag Switching is mainly a backbone technology, which is well suited for Internet Service Providers to efficiently route their Internet traffic across a high speed switching technology such as ATM.

Security and charging issues were not yet addressed in Tag Switching but they depend heavily on the used protocols.

Tag Switching is defined in a series of RFC and IETF drafts and is one of today's hottest topics in networking. Cisco announced the availability of a commercial implementation of Tag Switching by autumn 97.

6. Conclusion

In this paper we gave a technical overview on the competing integrated services network solutions, such as IP, ATM and the different available and emerging technologies on how to run IP over ATM networks, and identified their potential and shortcomings of being a solution for an integrated services network.

The following table summarizes some of the technical details of the discussed technologies.

Classification											
Pure IP Solution	x	x	x								
Pure ATM Solution				x					x		
IP over ATM Solution (overlay model)					x	x	x	x			
Label Switching Solution										x	x
Native ATM support						x	x	x	x	x	x
Emulation of LAN					x			x			
Multiple Layer 3 supported								x			x
Multiple Layer 2 supported	x	x	x				x				x
Network Scope											
Local Area Networks (LAN)	x	x	x	x	x	x	x	x	x		
Wide Area Networks (WAN)	x	x	x	x			x	x	x	x	x
Addressing											
IP addressing on application level	x	x	x		x	x	x	x		x	x
E.164/NSAP addressing on application level				x					x		
IP --> ATM address translation						x	x	x			
IP --> MAC address translation	x	x	x		x			x			
MAC --> ATM address translation					x			x			
User Data Encapsulation											
IP Packets	x	x	x		x	x	x	x	x	x	x
LLC/SNAP Packet encapsulation	x	x	x		x	x	x	x	x	x	x
MAC Packet encapsulation	x	x	x		x			x			
AAL5 Packet encapsulation				x	x	x	x	x	x	x	x
Data connection											
Connectionless	x	x	x							x	x
ATM permanent VCs				x	x	x		x		x	x
ATM switched VCs (Signalling used)				x	x	x	x	x	x		
End-to-end ATM connection in subnet				x	x	x	x	x	x	x	x
End-to-end ATM connection across subnets				x			x	x	x	x	x
Traffic Type											
Best Effort / UBR	x	x	x	x	x	x	x	x	x	x	x
Priority Based		x	x							x	x
ABR				x	x			x	x		
CBR				x		x	x	x	x		
VBR				x					x		

Table 1: Technological details

Table 2 can be used to compare the potential of the discussed technologies to satisfy the requirements of an integrated services technology.

Quality of Service Support											
Best effort delivery (also UBR)	x	x	x	x	x	x	x	x	x	x	x
QoS for IP flows in subnetwork			x	n/a	x	x	x	x	x		
QoS for IP flows across subnetwork			x	n/a			x	x	x		
QoS on application level in subnetwork			x	x							
QoS on application level across subnetworks			x	x							
Addressing Flexibility											
Unicast	x	x	x	x	x	x	x	x	x	x	x
Multicast	x	x	x	x				x		x	x
Anycast		x	x	x				x			
Deployment											
ATM Hardware				x	x	x	x	x	x	x	x
ATM Signalling Software				x	x	x	x	x	x		
ATM Routing Software				x	x	x	x	x	x		
IP Hardware	x	x	x		x	x	x	x		x	x
IP Signalling Software			x								
IP Routing Software	x	x	x		x	x	x	x		x	x
Integrates with legacy IP network equipment	x	x	x		x	x	x	x		x	x
Charging Support											
Flat rate network charging	x										
Traffic based network charging		x	x	x	x	x	x	x	x	x	x
Universal Connectivity											
Today	x	x	x		x	x	x	x		x	x
Future		x	x	x	x	x	x	x	x	x	x
Secuity Support											
Network Layer Security		x	x								
Availability Today											
Draft-Standards			x				x	x			x
Standards	x	x		x	x	x			x	x	x
Implementations available	x	x	x	x	x	x			x	x	
Applications available	x	x	x		x	x	x	x	x	x	x
						[1]	if IPv6 is used				
						[2]	not yet supported but possible				
						[3]	3 Q 97				

Table 2: Meeting the requirements of an integrated services network

It remains the question, which of these networking technologies will play a major role in the future. We try to answer this question for the short term and the medium-long term.

6.1 Short term

We think, that for the short term, IP is keeping its leading position on the network layer and ATM is only going to be used as a transport network because of the following reasons:

- There is a vast base of Internet equipment deployed in WAN and LAN area
- ATM deployment is scarce and mainly concentrated in the campus, backbone/WAN area
- There is an unmatched variety of applications based on IP
- There are hardly any applications available that make full use of the superior ATM features

This means, that for the short term the overlay model solutions like LANE and CLIP will play an important role in the deployment of ATM in LANs and backbones. Especially LANE's excellent potential of interconnecting and thus re-using legacy LAN equipment will make it the first choice of corporate network providers and ISPs who are willing to introduce ATM.

Using LANE or CLIP means, that there is no quality of service supported at the application layer, as the IPv4 layer hides all features of the underlying ATM network from higher layers. Only the high speed of ATM is exploited by these technologies.

If QoS support is requested by IP based applications today, a proprietary solution like Arequipa has to be used. Arequipa could play an important role in the short and medium term because it allows IP based application to request a quality of service by bypassing the IP layer while the necessary software only has to be deployed in the hosts and not in the entire network.

6.2 Medium-Long term

Despite the proceeding work of standardization organizations (ATM Forum, ITU, ETSI), it is not evident whether ATM will ever become an end-to-end solution because of various reasons:

- IP is extensively deployed (hardware and software)
- ATM software for signalling, routing, management and services is growing very complex and expensive
- Too much overhead to establish VC connections for short duration data flows
- Applications have to be changed considerably to use native ATM
- Full replacement of legacy LAN equipment needed to run end-to-end ATM

On the other hand, IETF's Integrated Services framework (IPv6, RSVP, ..) is catching up very fast with the ATM technology by introducing reservation, security and charging support and will therefore become a serious competitor for ATM.

This means that even in the future, several network technologies will coexist and ATM's role in this heterogeneous scenario will remain largely the one of a WAN/backbone technology instead of an end-to-end technology, for the reasons listed in the previous section.

ATM and IP will still coexist and solutions that can work in this heterogeneous world and that can take advantage of both technologies in a way that is transparent to the user, will play a key role. MPOA could potentially be used in the medium-long term, replacing LANE and CLIP which do not scale very well to large networks and can not offer end-to-end connections, assuming that the highly complex MPOA standard will ever be broadly accepted and implemented. It is more likely, that Label Switching technologies (i.e. IP Switching, Tag Switching) will be used in WANs/backbones because they can replace the huge control plane of ATM with the much simpler control software for label binding and label distribution. Using Label Switching technologies would reduce ATM to a mere transport technology.

As it is expected that the number of applications requesting QoS will increase, the demand of Integrated Services networks will raise, and ATM will have to co-operate to be able to provide end-to-end QoS in a heterogeneous network. The mapping of QoS requirements for different service classes between different technologies will be a key issue in the future.

7. Acknowledgments

One of the authors (S.G.), would like to thank Werner Almesberger (LRC-EPFL) Leena Chandran (LRC-EPFL) and Piergiorgio Cremonese for useful comments and suggestions.

References

[1] "Specification of Service Related Control Requirements",
ACTS AC094: EXPERT, WP2.1, Peter D. Soerensen et.al., Sept. 1996

[2] RFC 791: "Internet protocol"
J. Postel, September 1981

[3] RFC 1122: "Requirements for Internet hosts -- communication layers."
Braden, October 1989

[4] RFC 826: "Ethernet Address Resolution Protocol: Or converting network protocol address to 48 bit Ethernet address for transmission on Ethernet hardware"
D. Plummer, November 1982

[5] RFC 919: "Broadcasting Internet datagrams"
Mogul, October 1984

[6] RFC 1883: "Internet Protocol, Version 6 (IPv6) Specification"
Deering, R. Hinden, January 1996

[7] RFC 1884: "IP Version 6 Addressing Architecture"
Deering, R. Hinden, January 1996

[8] RFC 1885: "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6)"
A. Conta, S. Deering, January 1996

[9] RFC 1826: "IP Authentication Header"
R. Atkinson, August 1995

[10] RFC 1827: "IP Encapsulating Security Payload"
R. Atkinson, August 1995

[11] RFC 1933: "Transition Mechanisms for IPv6 Hosts and Routers"
R. Gilligan, E. Nordmark, April 1996

[12] IETF draft: "Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification"
Branden B. et all, October 1996.

[13] RSVP: A new Resource ReSerVation Protocol
Zhang, L., IEEE Network, September 1993.

[14] RFC 1112: "Host Extensions for IP Multicasting"
Deering, S., August 1989.

[15] IETF draft: "Building Blocks for Accounting and Access Control in RSVP"
Herzog, S., March 5, 1996

[16] "ATM User-Network Interface (UNI) Signalling Specification Version 4.0"
The ATM Forum, af-sig-0061.000

[17] "Specification of ATM Forum & ITU-T CS2.1 Signalling Functions",
ACTS AC094: EXPERT, WP2.1, Rolf M. Schmid et.al., March 1996

[18] "P-NNI V1.0"
The ATM Forum, af-pnni-0055.000

[19] "LAN Emulation over ATM 1.0"
The ATM Forum, af-lane-0021.000

[20] RFC 1577: "Classical IP and ARP over ATM"
Mark Laubach, January 1994

[21] RFC 1483: "Multiprotocol Encapsulation"
Mark Laubach, January 1994

[22] RFC 1755: "ATM Signalling Support for IP over ATM"
M. Perez, F.A. Mankin, E. Hoffman, G. Grosman, A.Malis, February 1995

[23] RFC 1293: "Inverse Address Resolution Protocol"
T. Bradley, C. Brown, January 1992

[24] IETF draft: "Multicast Address Resolution Server (MARS)"
draft-ietf-ipatm-ipmc-12.txt

[25] RFC 2170: "IP over ATM (Arequipa)"
Almesberger, Le Boudec, Oechslin, January 1996

[26] RFC 1953: "Ipsilon Flow Management Protocol Specification fo IPv4, version 1.0"
P.W. Edwards, R.E. Hoffman, F. Liaw, T. Lycon, G. Minshall, May 1996

[27] RFC 1954: "Transmission of Flow Labelled IPv4 on ATM Data Links, Ipsilon Version 1.0"
P. Newman, W. Edwards, R. Hinden, E. Hoffman, F. Liaw, May 1996

[28] RFC 1987: "Ipsilon's General Switch Management Protocol Specification Version 1.1"
P. Newman, W. Edwards, R. Hinden, E. Hoffman, F. Liaw, August 1996

[29] IETF draft: "NMBA Next Hop Resolution Protocol (NHRP)"
draft-ietf-rolc-nhrp-010.txt.
J. Luciani, D. Katz, D. Piscitello and B. Cole,

[30] RFC 1209: "Transmission of IP datagrams over the SMDS service"
J. Lawrance and D. Piscitello, March 1991

[31] RFC 1937: "Local/Remote" Forwarding Decision in Switched Data Link Subnetworks"
Yakov Rekhter and Dilip Kandlur

[32] MPOA Baseline Version 1, ATM Forum Contribution,
ATM Forum Sub-Working Group, September 1996

[33] IETF draft: "NHRP for Destinations off the NBMA Subnetwork"
draft-ietf-rolc-r2r-nhrp-00.txt.
Y. Rekhter

[34] "Tag Switching Overview"
Berkeley EECS Seminar, Y. Rekhter, May 1997

[35] RFC 2105: "Cisco Systems' Tag Switching Architecture Overview"
Y. Rekhter, B. Davie, D. Katz, E. Rosen, G. Swallow, February 1997

[36] IETF draft: "Tag Distribution Protocol"
draft-doolan-tdp-spec-01.txt
P. Doolan, B. Davie, D. Katz, Y. Rekhter, E. Rosen, May 1997

[37] IETF draft: “Use of Tag Switching With ATM”
draft-davie-tag-switching-atm-01.txt
P. Davie, P. Doolan, J. Lawrence, K. McCloghrie, Y. Rekhter, E. Rosen, G. Swallow, January 1997

[38] RFC 1122: “Requirements for Internet Hosts”
R. Braden, 1989

Abbreviations

AAL5	ATM Adaptation Layer 5
ABR	Available Bit Rate
ACTS	Advanced Communications Technologies and Services
API	Application Programmer Interface
ARIS	Aggregate Route-Based IP Switch
ARP	Address Resolution Protocol
Arequipa	Application REQuested IP over ATM
ATM	Asynchronous Transfer Mode
BGP	Border Gateway Protocol
B-ICI	Broadband Inter Carrier Interface
B-ISUP	Broadband ISUP
BLLI	Broadband Low Layer Information
BUS	Broadcast and Unknown Server
CBR	Constant Bit Rate
CIDR	Classless Inter-Domain Routing
CLIP	CLassical IP over ATM
CSR	Cell Switched Router
DFFG	Default Forwarder Functional Group
DNS	Domain Name System
DS-1	Digital Signal Level 1
DS-3	Digital Signal Level 3
EGP	Exterior Gateway Protocol
ELAN	Emulated LAN
EPFL	Ecole Polytechnique Federale de Lausanne
ESP	Encapsulating Security Payload
GSMP	General Switch Management Protocol
HTML	Hypertext Markup Language
IASG	Inter Address Sub-Group
ICFG	IASG Coordination Functional Group
ICMP	Internet Control Message Protocol
IEEE	Institute of Electrical and Electronic Engineers
IETF	Internet Engineering Task Force
IFMP	Ipsilon Flow Management Protocol
IGP	Interior Gateway Protocol
IHL	Internet Header Length
InARP	Inverse ARP
ION	IP over NBMA
IP	Internet Protocol
IPng	IP next generation
IPv4	IP version 4
IPv6	IP version 6 (=IPng)
ISP	Internet Service Provider
ITU	International Telecommunication Union
LAG	Local Address Group
LAN	Local Area Network
LANE	LAN Emulation
LE_ARP	LAN Emulation ARP
LEC	LAN Emulation Client
LECS	LAN Emulation Configuration Server
LEC	LAN Emulation Client
LIS	Logical IP Subnetwork
LLC	Logical Link Control
MAC	Media Access Control
MARS	Multicast Address Resolution Server
MCR	Minimum Cell Rate
MPOA	Multi Protocol Over ATM
MPLS	Multiprotocol Label Switching
MSF	Multicast Server Function
MTU	Maximum Transfer Unit
N-ISDN	Narrowband ISDN
NBMA	Non-Broadcast Multiple-Access
NHRP	Next Hop Resolution Protocol
NHC	Next Hop Clients
NHS	Next Hop Server
NIC	Network Interface Card
NNI	Network-Node Interface
NSAP	Network Service Access Point
IGMP	Internet Group Management Protocol
ISDN	Integrated Services Digital Network
ISUP	Integrated Services User Part
IPX	Internetwork Packet eXchange
OS	Operating System
OSPF	Open Shortest Path First
PCR	Peak Cell Rate
PNNI	Private NNI
PNO	Public Network Operator
POTS	Plain Old Telephone System
PVC	Permanent VC
QoS	Quality of Service
RARP	Reverse ARP
RFC	Request For Comments
RFFG	Remote Forwarder Functional Group
RSFG	Route Server Functional Group
RSVP	Resource reSerVation Protocol
SAR	Segmentation and Reassembly
SCR	Sustained Cell Rate
SITA	Switching IP Through ATM
SMDS	Switched Multimegabit Data Service
SNAP	Sub Network Access Point
SVC	Switched VC
TDP	Tag Distribution Protocol
TIB	Tag Information Base
TCP	Transmission Control Protocol
TOS	Type Of Service
UBR	Unspecified Bit Rate
UDP	User Datagram Protocol
UNI	User-Network Interface
VBR	Variable Bit Rate
VC	Virtual Connection
VCC	Virtual Channel Connection
VCI	Virtual Channel Identifier
VLAN	Virtual LAN
VoD	Video on Demand
VPI	Virtual Path Identifier
WAN	Wide Area Network
WWW	World Wide Web