Andrew G. Malis. “The Development Of Multiprotocol Label Switching”
Handbook of Emerging Communications Technologies: The Next Decade.
Ed. Saba Zamir
Boca Raton: CRC Press LLC, 2000

8 The Development Of Multiprotocol Label Switching

To Integrate IP With ATM For The Internet Backbone

Andrew G. Malis

CONTENTS

8.1 INTRODUCTION

This chapter describes the many methods developed to integrate IP routing with ATM switching, culminating in the development of MPLS, or Multiprotocol Label Switching. MPLS is a set of standards being developed by the Internet Engineering Task Force (IETF) to allow Layer 3 IP forwarding (traditionally performed by routers) to be combined with Layer 2 switching (such as ATM and Frame Relay) in

© 2000 by CRC Press LLC

a single combined system, to increase the scalability, speed, and traffic engineering capabilities of the Internet. It also adds additional functionality in routers such as traffic engineering (also known as service provisioning and bandwidth management) and virtual private networks that has traditionally been found only in switches.

This chapter provides an introduction to MPLS technology by describing the numerous methods of integrating IP and ATM that led to MPLS development. Sections 8.2 and 8.3 describe the overlay model of running IP on top of ATM which is used by classical IP over ATM, NHRP, MPOA, and simple interconnection of routers using ATM permanent virtual circuits. Section 8.4 describes the integrated model of combining IP and ATM, of which MPLS is one example, and Section 8.4.5 discusses MPLS in particular.

8.2 EVOLUTION OF THE INTERNET BACKBONE

The need for MPLS evolved from the rapid growth of the Internet and the inability of IP routers to keep up with the raw bandwidth requirements of the Internet backbone. Traditional router architectures included both routing (using a distributed algorithm to determine the path to a particular destination IP address) and forwarding (using the output of the routing algorithm to choose the output interface for a particular IP packet) as software tasks in a shared processor. Often forwarding would be a high-priority foreground task and routing a lower priority background task. More recent router architectures have improved their performance by moving the forwarding tasks to dedicated processors; often the number of forwarding processors increases linearly with the number of interfaces on the router. Most recently, gigabit routers have begun to move the forwarding task from general purpose processors to specialized application-specific integrated circuits (ASICs). Note that the routing protocols continue to run as a software background task in a general purpose processor; in almost all cases, this is the same processor also used for other background tasks, such as network management activities.

However, router forwarding speeds have still not been able to keep up with the raw bandwidth requirements of the Internet. In addition, the variable-sized packet nature of IP and the datagram hop-by-hop nature of IP routing protocols make it extremely difficult to support advanced features such as Quality of Service (QoS) and traffic engineering.

Consequently, a majority of the Internet backbone carriers have chosen to use an ATM-based network backbone. Using ATM to interconnect the routers at the edge of the carrier network has a number of advantages:

- Each ATM cell's short, fixed-length label, known as the virtual path identifier/virtual circuit identifier (VPI/VCI), is very simple to switch in hardware. IP's longer and hierarchically-structured addresses are much harder to switch. ATM switches generally switch cells at the line rate of the interfaces; routers often cannot keep up with the full rate of their input interfaces.

© 2000 by CRC Press LLC

- The uniform cell length, 53 octets (including the five-octet cell header), simplifies buffering and queue management algorithms in the switch when compared to a router.
- The uniform cell length also makes it possible to support sophisticated QoS services in the network, especially if the same backbone is used for other services in addition to IP (such as providing native ATM services, Frame Relay, private line emulation, and so on).
- Because ATM uses virtual circuits rather than datagram forwarding, the routes that cells take through the ATM network are independent of the IP routing protocol, allowing network administrators complete control over traffic routing and the loads on network trunks.

However, because IP and ATM are so different, adapting IP to ATM networks has been easier said than done, and many approaches have developed.

8.3 INTEGRATING IP AND ATM: THE OVERLAY MODEL

8.3.1 Multiprotocol Encapsulation

Before IP was able to be run within ATM circuits, a number of fundamental issues needed to be settled because of the basic differences between IP's connectionless datagram packet nature and ATM's connection-oriented cell nature. The first issue was basic encapsulation – how should variable-length IP packets be carried in ATM cells, and how should the higher-layer (with respect to ATM) protocol be identified, in order to be able to multiplex multiple higher-layer protocols over a single ATM virtual circuit (VC)?

The IETF, which produces all IP-related protocol standards, defined a method to encapsulate IP in ATM and to multiplex multiple protocols on the same ATM VC in RFC (Request for Comments) 1483.[1] For encapsulation, the IETF chose to use *ATM Adaptation Layer 5* (AAL5).[2] The decision of which protocol identification method to use was not as straightforward as it might seem because there were several among which to choose. WAN networking protocols, such as X.25 and Frame Relay, already used single-octet Network Layer Protocol IDs (NLPIDs) defined in ISO/IEC Technical Report 9577.[3] Some of the more popular NLPIDs are shown in Table 8.1.

TABLE 8.1
Selected NLPIDs

0xCC:	IP
0x81:	CLNP
0x08:	Q.933
0x80:	SNAP (see below)

© 2000 by CRC Press LLC

Meanwhile, LANs used logical link control (LLC) and subnetwork attachment point (SNAP), as defined by the IEEE for protocol identification on 802 LANs such as 802.3 Ethernet and 802.5 Token Ring.[4] LLC uses three octets to define the following protocol in the LAN frame. These three octets are most often 0xAA-AA-03, which signify SNAP, which occupies another five octets. SNAP has two fields: a three-octet organizationally unique identifier (OUI) and a two-octet protocol identifier (PID). The meaning of the PID field is dependent on the value of the OUI field; this allows different organizations to define their own sets of protocols to be carried over LANs. LLC and SNAP are used together so often that they are usually referred to as LLC/SNAP. Interestingly, SNAP's use is not restricted to LLC; there is a NLPID, 0x80, which signifies that it is immediately followed by a SNAP header. This extends NLPIDs to be able to identify any protocol identified by SNAP.

The list of SNAP OUIs is administered by the IEEE, and any organization, company, or individual can be granted an OUI by paying a registration fee to the IEEE. However, the most complete list of OUIs is not published by the IEEE, but by the IETF's Internet Assigned Numbers Authority (IANA) in its document called "Assigned Numbers," which is periodically reissued as a new RFC. As of this writing, the most recent version of "Assigned Numbers" is RFC 1700,[5] but its most up-to-date contents can be found on the Internet at ftp://ftp.isi.edu/in-notes/iana/assignments/.

Some example OUIs are provided in Table 8.2.

TABLE 8.2
Selected OUIs

0x00-00-00	Xerox (see below)
0x00-00-5E	IANA
0x00-80-C2	IEEE 802.1 Committee
0x00-A0-3E	ATM Forum
0x02-60-8C	3Com (an example corporate assignment)

One of the OUIs is special: 0x00-00-00 is assigned to and administered by Xerox (one of the inventors of Ethernet), and it is used to record the list of protocol IDs used on the original Ethernet, which uses a simple two-octet protocol identifier rather than the eight-octet LLC/SNAP. As a result, most of the major multivendor internetworking protocols can be found in this OUI, including IP, IPX, Appletalk, etc. The list of protocol IDs in OUI 00-00-00 are also known as *Ethertypes* (after the original name of the Ethernet Type field), and new Ethertypes can be obtained from Xerox for a fee. Again, "Assigned Numbers" contains an extensive list of assigned Ethertypes.

Some of the more popular Ethertypes are listed in Table 8.3

After a considerable amount of debate, the IETF chose in RFC 1483 to use LLC/SNAP for multiprotocol identification. Their primary reason for this decision was that LLC/SNAP's fixed size of eight octets allowed more efficient processing

© 2000 by CRC Press LLC

TABLE 8.3
Selected Ethertypes

0x08-00	IP
0x08-06	ARP
0x80-9B	Appletalk
0x81-37	IPX

than variable length (but more compact) NLPIDs. The IETF also had a greater degree of familiarity with LLC/SNAP. Of course, using NLPIDs would have eased ATM/Frame Relay interoperation, but this consideration was not given much weight. The one exception when NLPIDs are used for protocol identification over ATM is when Frame Relay is being layered over ATM via the frame relay service specific convergence sublayer (FR-SSCS), as specified in Appendix A of RFC 1483.

8.3.2 IP Over ATM Using Permanent Virtual Connections (PVCs)

By far, the most popular method of carrying IP over ATM in the WANs that make up the Internet backbone is by interconnecting the routers at the edge of the WAN with a full mesh of ATM PVCs interconnecting the routers. Each ATM PVC emulates a point-to-point interconnection between the routers. Over each of the PVCs, RFC 1483 is used to encapsulate the IP packets in ATM cells and identify IP as the protocol being carried. In particular, the OUI 0x00-00-00 and the PID 0x08-00 are used to identify IP. Because a full mesh of PVCs is used, the physical topology of the network at the ATM layer and the logical topology as seen by the routers are quite different. Figure 8.1 shows an example physical topology of routers and ATM switches.

Note that the routers and ATM switches are numbered differently, to emphasize their independence from each other, and that the ATM switches are typically interconnected with a partial mesh of switch-switch trunks. The ATM network has a complete set of ATM PVCs configured between the routers, which produces the logical topology shown in Figure 8.2.

Because the logical topology is overlaid onto the physical ATM network, with separate addressing plans and topologies, this sort of arrangement is known as the overlay model. This model requires N*(N-1) PVCs to interconnect N routers, which does not scale well to large networks. It also produces a large number of routing adjacencies at the IP routing layer, which increases the overhead required to run the IP routing protocols. However, it does have one particular advantage, which is evident when the PVCs are shown overlaying the ATM network as in Figure 8.3.

Because the topologies are independent, the network operator can choose how to route the PVCs through the physical network. For example, the PVC between routers 3 and 5 can be routed at the ATM layer between S5, S6, and S8, or between S5, S7, and S8, whichever is more useful for producing the desired traffic loading on the ATM trunks. Network operators have found this ability to be extremely useful.

© 2000 by CRC Press LLC

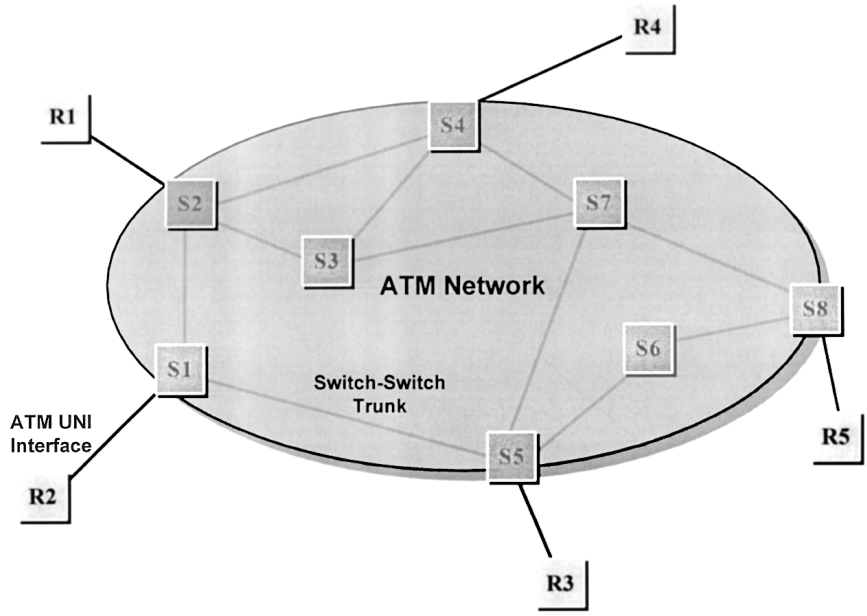

Figure 8.1 ATM-Connected Routers: Physical Topology

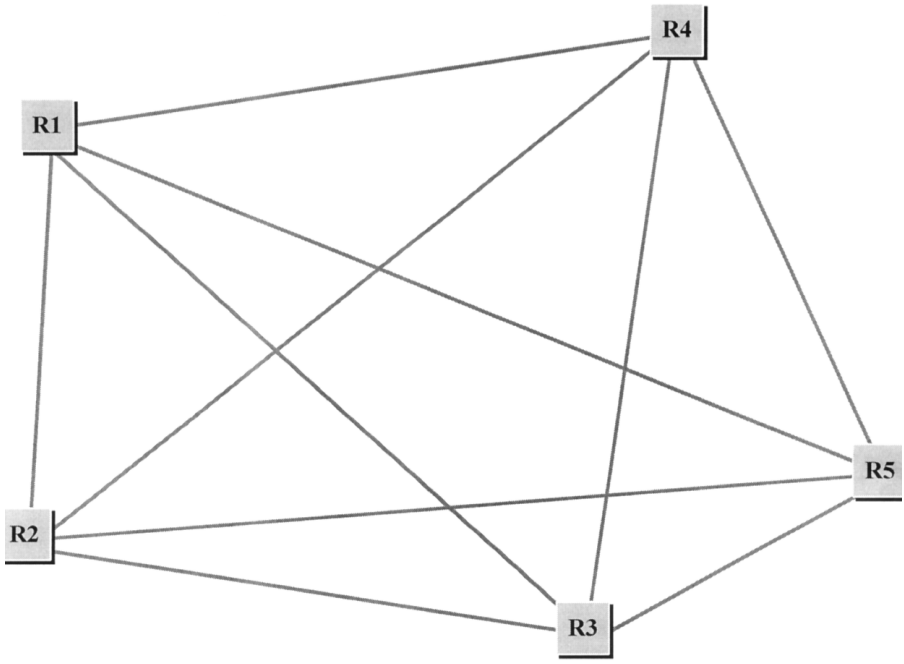

Figure 8.2 ATM-Connected Routers: Logical Topology

© 2000 by CRC Press LLC

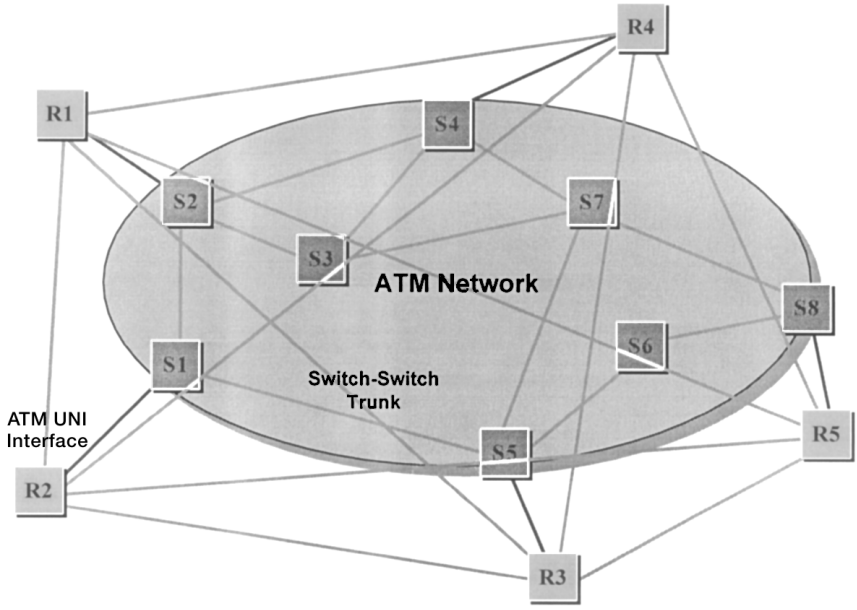

Figure 8.3 ATM-Connected Routers: Both Topologies

8.3.3 IP Over ATM Using Switched Virtual Circuits (SVCs)

8.3.3.1 Classical IP Over ATM and ATMARP

When there are too many ATM-attached IP end systems (whether routers or ATM-attached hosts, such as workstations) for a full PVC mesh to be practical, as might be found on an ATM LAN or in an enterprise ATM network, the overlay model can still be used, but with ATM SVCs rather than PVCs. The SVCs can be created when they are required to interconnect two end systems then destroyed when they are no longer needed. The IETF's "*Classical IP over ATM*" set of standards (RFCs 1755,[6] 2225,[7] and 2331[8]) form the IETF-approved method of carrying IP over ATM SVCs.

Classical IP over ATM is so named because it is the closest possible adaptation of the any-to-any connectivity provided by "classical" IP (such as used on a LAN) such that it would work over ATM's connection-oriented and nonbroadcast nature. As with PVCs, classical IP over ATM overlays IP over ATM; there is no direct relationship between the IP addresses in the IP packets and the ATM addresses used to open VCs to carry the IP packets.

There are three RFCs that together describe classical IP over ATM:

RFC 2225: How IP addresses are resolved to ATM addresses, and subsequently transported over ATM connections.

© 2000 by CRC Press LLC

RFC 1755: How to use ATM Forum UNI 3.0[9] and 3.1[10] SVC signaling when opening RFC 1577 connections.

RFC 2331: Updates to RFC 1755 to support ATM Forum UNI 4.0[11] SVC signaling.

Because ATM networks can support many more stations (hosts and routers) than are found on typical LANs, a single ATM network can have more than one IP subnetwork layered on top of it, rather than using the typical strategy of having a one-to-one correspondence between the IP subnetwork and the physical network over which it is layered. Each of these subnetworks is called a logical IP subnetwork (LIS).

Figure 8.4 shows an ATM network with two IP LIS layered above it. LIS are purely logical entities that have been layered above the physical ATM network, and LIS membership is also purely logical (any host on the ATM network can belong to any LIS layered above it). Two IP stations in the same LIS communicate directly via ATM SVCs or PVCs, while IP stations in different LIS must intercommunicate via a router (such as R1). Multiple LIS are used for administrative requirements, such as security filtering, convenience, or to localize dependencies on (and reduce the load of) ATMARP and other network servers.

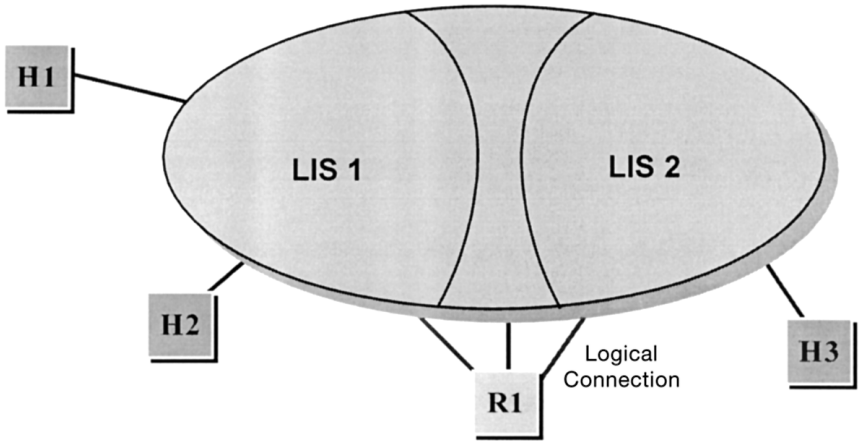

Figure 8.4 Two IP LIS on One ATM Network

Router R1, while having only one physical attachment to the ATM network, has been defined to logically belong to both LIS so that it may route between them. Because hosts H1 and H2 are in the same LIS, they intercommunicate directly by opening an ATM SVC from one to the other when there is data to send (see Figure 8.5).

Host H3, however, must indirectly communicate with hosts H1 and H2 by opening an SVC to router R1, which will in turn open an SVC to host H1 and/or H2 as required, as shown in Figure 8.6.

© 2000 by CRC Press LLC

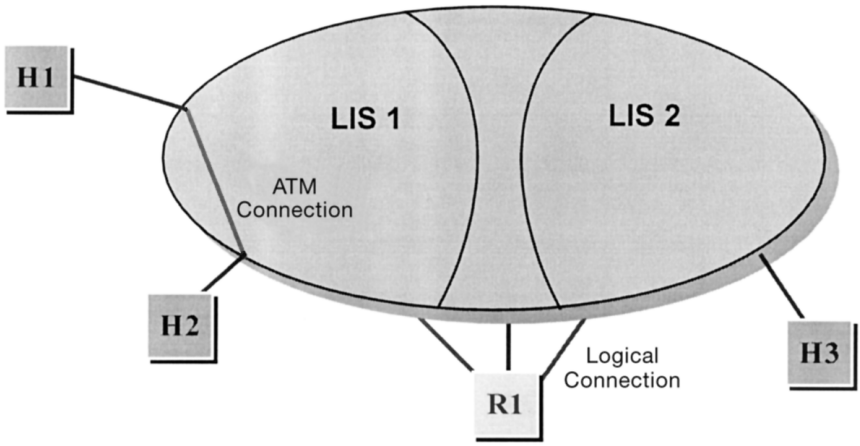

Figure 8.5 Connection between Hosts H1 and H2

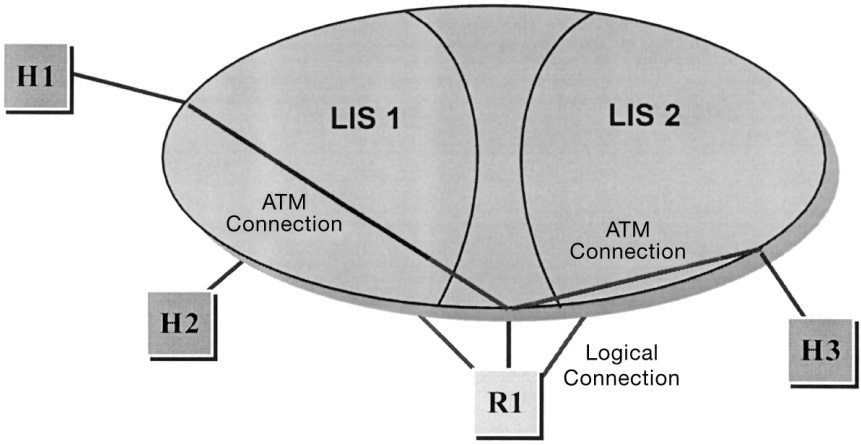

Figure 8.6 Connection between Hosts H1 and H3

Because there is no direct relationship between IP and ATM addresses, hosts need some method to resolve IP addresses to ATM addresses, in order to open SVCs to each other so they can communicate at the ATM layer when they have IP packets to send. RFC 2225 defines an ATMARP server and protocol to be used for such address resolution. This is necessary because IP over Ethernet (and other LANs) is able to broadcast ARP (Address Resolution Protocol) requests to all of the other stations on the LAN; however, because ATM does not have a broadcast mechanism, a server must be used instead. Each LIS requires an ATMARP server to translate IP to ATM addresses.

© 2000 by CRC Press LLC

ATMARP operates as follows: When hosts H1 and H2 come up, they register themselves with their LIS' ATMARP server ("AS" in Figure 8.7) by opening connections to a configured ATMARP server, and issuing an ATMARP registration. Note that while the figure shows a stand-alone ATMARP server, they are typically implemented as a logical function in an ATM switch or an ATM-attached router, rather than as a separate piece of equipment.

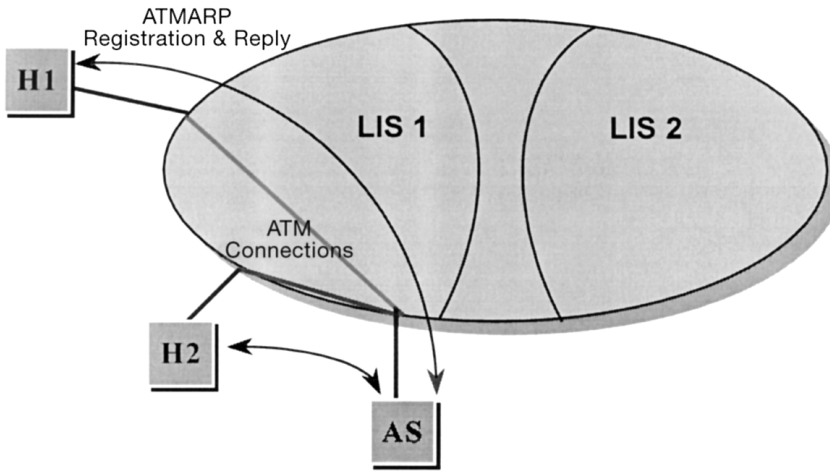

Figure 8.7 Connections to the ATMARP Server

The ATMARP server now contains the mapping between host H1 and H2's IP and ATM addresses. As shown in Figure 8.8, when host H1 needs to send IP packets to host H2, it sends an ATMARP request containing host H2's IP address to the ATMARP server. The server will return host H2's ATM address to host H1, allowing host H1 to open a direct ATM connection to host H2.

Figure 8.9 shows the resulting connection to host H2 using its ATM address returned by the server:

8.3.3.2 Next Hop Resolution Protocol (NHRP)

As Figure 8.6 showed, only hosts in the same LIS may open direct ATM connections to each other; hosts in separate LIS must communicate via a router, even if it means the packets must leave the ATM network only to reenter the ATM network on the same physical access line to the router (but on a different VC, of course).

Figure 8.10 shows a more complicated example. The ATM network contains three LIS, and host H1 on LIS 1 needs to communicate with host H2 on LIS 3. Router R1 is logically on both LIS 1 and 2, and router R2 logically interconnects LIS 2 and 3. Because the hosts are in different LISes, they must indirectly communicate along the path provided by IP routing: from host H1 to router R1, then to router R2 on a different VC, and then to host H2 on a third VC.

© 2000 by CRC Press LLC

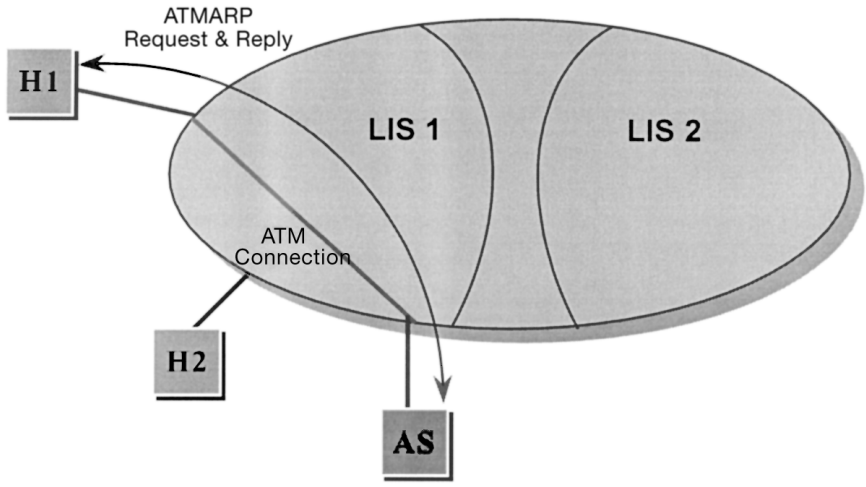

Figure 8.8 Host H1 Requesting Host H2's ATM Address

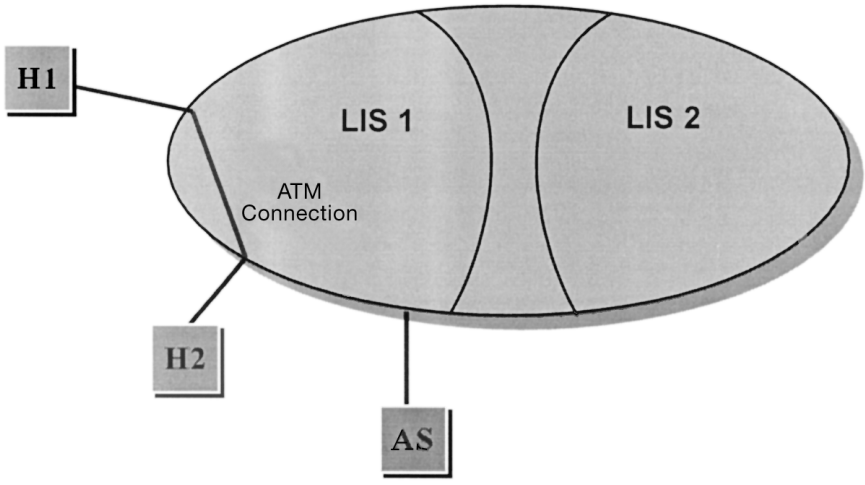

Figure 8.9 Resulting Connection Between H1 and H2

At the routers, the IP packets leave and reenter the ATM network. This is extremely inefficient, in that the packets must traverse each router's interface line twice, their cells must be reassembled into packets as they enter the router, each packet must be processed and switched by the router, and they must be resegmented into cells as they reenter the ATM network using the same ATM access line.

To prevent this inefficiency, the IETF developed the next hop routing protocol (NHRP)[12] to allow IP hosts in different LIS, but on the same ATM network, to be

© 2000 by CRC Press LLC

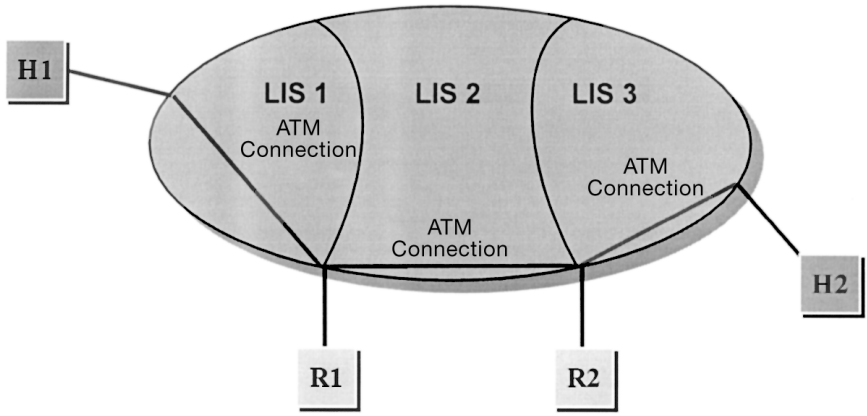

Figure 8.10 Routed Path From A to B

able to open direct VCs between each other (unless there are administrative reasons for this to be prevented).

NHRP can be viewed as an enhancement of ATMARP; ATMARP clients (end stations) become NHRP clients, and ATMARP servers are replaced by NHRP Servers (NHS). Because the client behavior is very similar to that of ATMARP, upgrading the client software is simple. Because NHRP is closely tied to IP routing, NHS are almost always implemented in IP routers (or ATM switches that include IP routing functionality).

When hosts H1 and H2 come up, they issue an NHRP registration request to their local NHS. In most cases, their default router will also be their NHS, so no extra configuration is required in the hosts. In Figure 8.11, router R1 is also the NHS for LIS 1, and router R2 is the NHS for LIS 3. While it is not germane to this example, LIS 2 could be served by either (or both) of the routers, or by a third router.

When host H1 needs to communicate with host H2, and it knows only H2's IP address, it issues an NHRP resolution request for H2's ATM address. Note that hosts H1 and H2 no longer need to be in the same LIS; host H1 can issue an NHRP resolution request for all hosts with which it needs to communicate.

As shown in Figure 8.12, if host H1 issues an NHRP resolution request to its server (router R1), and its server does not know the mapping between host H2's IP and ATM addresses, it forwards the request along the normal IP routed path towards host H2, until the request arrives at a router/NHS that knows the mapping (in this case, router R2). Note that the request stops at the first server that knows the answer, and is never sent to host B itself. (Keep in mind that ATM SVCs need to be opened to carry these signaling messages if there is not already a connection open.)

© 2000 by CRC Press LLC

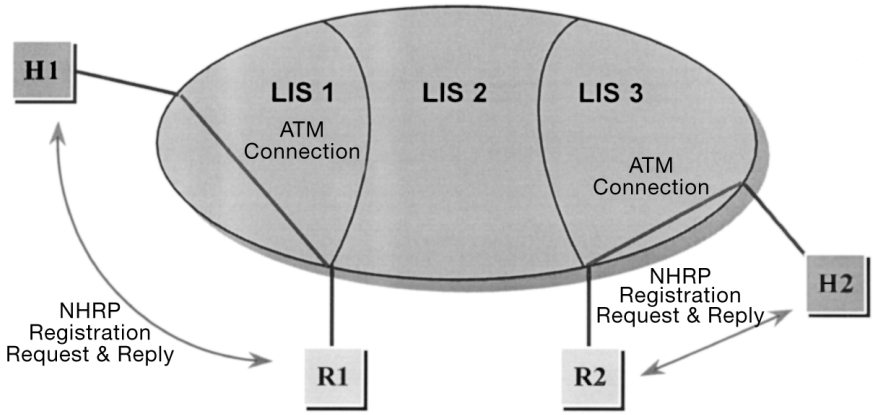

Figure 8.11 NHRP Registration

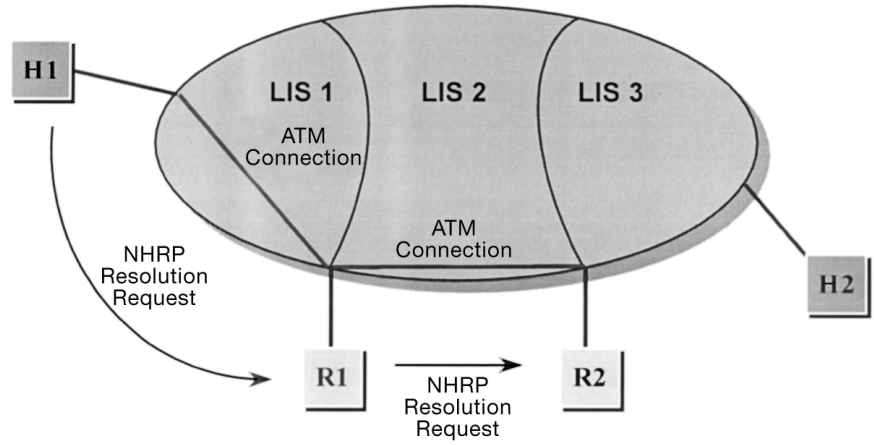

Figure 8.12 NHRP Query Processing

In Figure 8.13, the NHRP resolution reply is returned back along the same path to host H1. The answer is also cached by any NHSes on the way (router R1 in this example), so that future requests can be answered locally.

Finally, host H1 is able to open a direct ATM connection to host H2, bypassing the routers, as shown in Figure 8.14.

© 2000 by CRC Press LLC

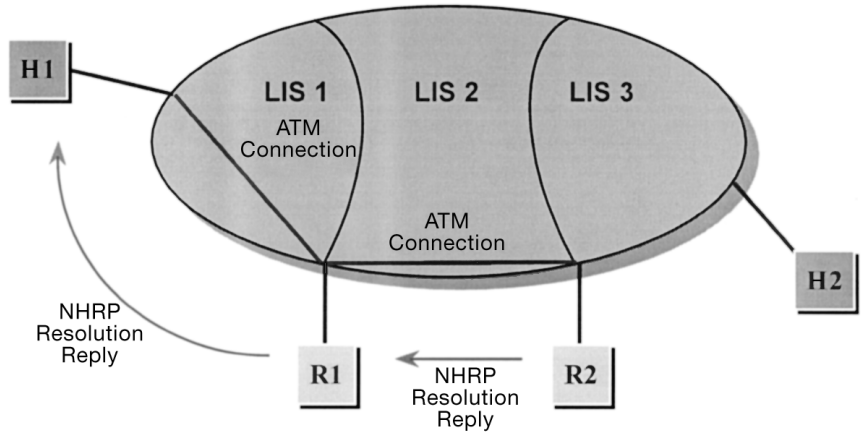

Figure 8.13 NHRP Resolution Reply

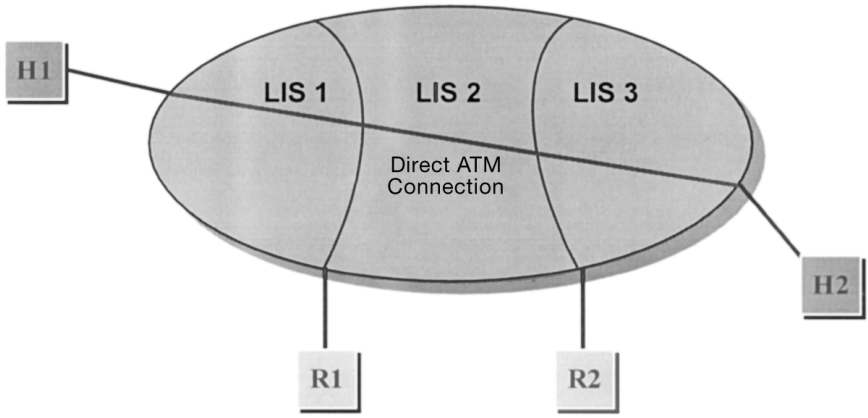

Figure 8.14 Direct Connection Between A and B

8.3.3.3 Multiprotocol Over ATM (MPOA)

The ATM Forum used NHRP as the basis for its multiprotocol over ATM (MPOA) [13] specification. MPOA, like Classical IP over ATM and NHRP, is based on the overlay model and relies on an NHRP server to provide address resolution services. In addition, it uses ATM Forum LAN Emulation (LANE)[14] to provide both bridging and routing support to MPOA clients. However, this functionality exacts a price; it uses multiple additional servers, including an MPOA server and several LANE servers and it requires ATM-level connections between the clients and multiple servers.

© 2000 by CRC Press LLC

Classical IP over ATM, NHRP, and MPOA all work well in enterprise or local-area ATM networks. However, they all share serious limitations that do not allow them to scale for WAN use, such as on an Internet backbone network:

- They all open ATM connections to carry data based on detecting new IP traffic flows. This can lead to delays while connections open. In addition, many ATM switches and ATM router interfaces are seriously limited to the speed at which they can open ATM connections and maximum number of connections they can have open simultaneously. This is an extreme restriction because Internet backbone networks routinely carry 250,000 simultaneous flows.[15]
- There is excessive management complexity and overhead as a result of the large number of servers and the necessity to run both IP and ATM routing protocols.
- The lack of server redundancy in the LANE and MPOA specifications makes these protocols difficult to use as the basis of a public service network.

All of these restrictions led to the development of the integrated model of IP and ATM, which is discussed in the next section.

8.4 INTEGRATING IP AND ATM: THE INTEGRATED MODEL

The scaling limitation inherent in the overlay model of IP over ATM, whether for PVCs or SVCs, led to the development of a number of proprietary solutions by a number of IP router and ATM switch vendors. They are discussed in the following sections in the approximate order of their development. They have culminated in the development of MPLS as a multivendor, interoperable solution for integrating IP and ATM in the wide area.

8.4.1 IP Switching

IP switching is a term invented by Ipsilon Networks, Inc. (which has since been acquired by Nokia Telecommunications, Inc.). Rather than using the classical IP over ATM approach of integrating IP and ATM, they chose to pair every ATM switch with a router that ran a new protocol, the Ipsilon Flow Management Protocol (IFMP).[16] IFMP was an alternate ATM SVC signaling protocol, which, like Classical IP, NHRP, and MPOA, would detect IP flows and open ATM connections to carry them. Once an IP flow was mapped to an ATM connection, it would be switched through the network at the ATM layer, rather than routed through the network at the IP layer. Ipsilon's main innovation was replacing the standard ITU and ATM Forum SVC signaling protocols with a much simpler protocol (IFMP), which allowed connections to complete faster and was much easier for them to implement. Also,

© 2000 by CRC Press LLC

not all flows were mapped to ATM connections; low bandwidth traffic continued to be routed over a default ATM connection over the trunk between two router/switches.

However, Ipsilon's IP switching had several problems that restricted its suitability for use in wide-area Internet backbone networks:

- Because it opened ATM connections as a result of observing IP traffic flows, there was a delay that allowed packets to get out of order as they moved from the default ATM connection to a dedicated connection. Packet misordering slows down many TCP implementations.
- The switches were limited to the maximum number of flows they could support simultaneously. This problem was made worse by their flow detection algorithm, which was very fine-grained; it would open a connection based now on only the source and destination IP addresses, but also the particular application that was in use. This could result in multiple ATM connections to carry flows between the same IP source and destination hosts.
- Ipsilon refused to submit IFMP for multivendor standardization, preferring to publish it as a proprietary protocol. Most wide-area service providers prefer to use standardized protocols so they are not locked into a single vendor's products.

8.4.2 IP Navigator

IP Navigator,[17] developed by Cascade Communications Corporation (which was later acquired by Ascend Communications, Inc.), was the first method of integrating IP and ATM that based the ATM connections on IP routing. Cascade observed that in Internet backbone networks, IP routing is generally stable – it changes only during outages or as new IP networks are added to the Internet. In addition, they realized it would be easiest to interconnect to other parts of the Internet if no new protocols were required in external routers. Therefore they created IP Navigator, which has the following functionality:

- A network of Frame Relay and/or ATM switches running IP Navigator appear as a collection of routers, not switches, to outside routers. IP Navigator supports standard IP routing protocols, such as OSPF,[18] BGP, and so on. It presents standard IP router interfaces to its neighboring routers.
- Inside the network, the switches use OSPF for two functions. The first is to perform the standard IP routing functionality. The second is to use the switch-switch paths determined by OSPF's IP routing to set up switched connections at the Frame Relay or ATM layer. Because the connections are formed as a result of a network control protocol (OSPF) rather than actual traffic flows, IP Navigator is said to be *control-driven* rather than *flow-driven*, like IFMP or MPOA.
- These control-driven connections are used to carry IP traffic from the network ingress to egress. When an IP packet enters the network, it is

© 2000 by CRC Press LLC

routed by the ingress switch, using a standard IP routing table, onto the proper Frame Relay or ATM connection to switch the packet to the correct egress switch. The packet is switched through the interior of the network, then routed again at the egress switch to find the proper egress interface.

The key to IP Navigator's scalability for the wide area is that it does not maintain a full mesh of connections from every ingress switch to every egress switch (after all, every switch with external interfaces is both an ingress and an egress). Rather, it uses point-to-multipoint trees (MPT) to aggregate the packets from the ingress switches to the proper egress switch.

An MPT is the opposite of a Frame Relay or ATM point-to-multipoint connection. Rather than traffic being sent from the root to the leaves of the connection, the IP packets are sent from the leaves of the tree to the root. There is one tree, and one root, for every egress switch. This allows the number of connections to scale with the number of switches, rather than square of the number of switches.

When the IP Navigator network is initialized, standard OSPF routing is used to determine the path from every ingress switch to each egress switch. IP Navigator then uses that information to create an MPT for every egress. This happens automatically, and prior to any user traffic being sent. If IP routing should later change, perhaps because of a power outage at a switch, then the network will automatically reform the MPTs to match. In each interior switch, packets are switched by their MPT.

Figures 8.15 and 8.16 show an example MPT. Figure 8.15 shows the physical network topology. As far as the external routers (R1-R5) are concerned, switches S1-S8 are actually routers and running standard routing protocols.

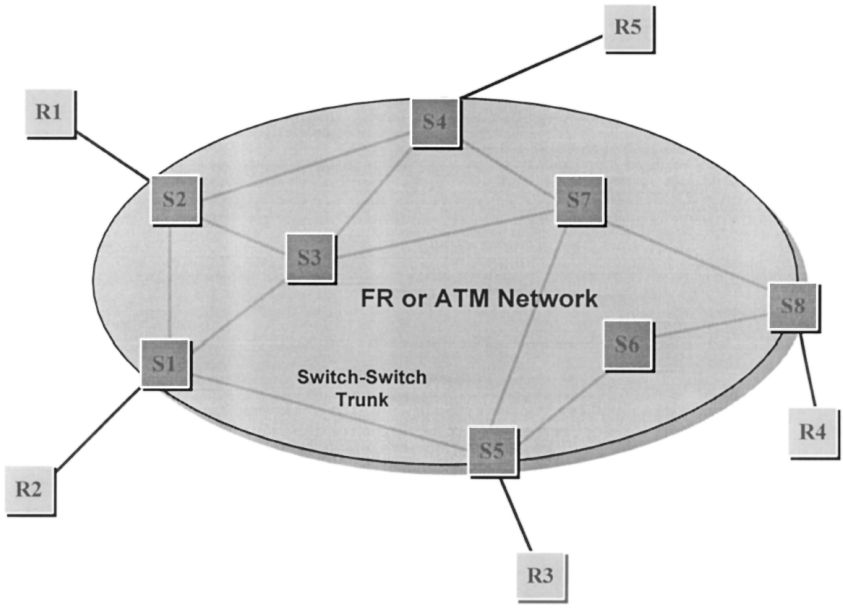

Figure 8.15 Physical Network Topology

© 2000 by CRC Press LLC

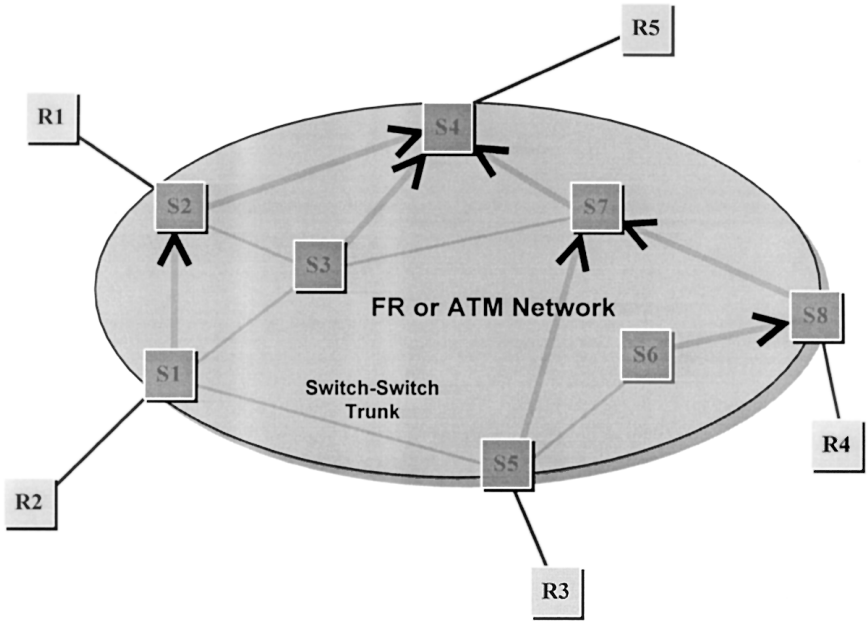

Figure 8.16 IP Navigator MPT with S4 as the Egress

Internally, however, the switches have set up MPTs to carry the IP traffic so that the packets do not have to be routed at each hop through the network.

Figure 8.16 shows an MPT with S4 as the egress switch.

In this example, IP packets from R2 to R5 will travel through switches S1, S2, and S4, while packets from R3 to R5 will travel though switches S5, S7, and S4. Note that S7 aggregates packets from S5, S6, S7, and S8 onto the MPT to S4.

IP Navigator implements MPTs over Frame Relay by assigning each egress switch a unique Frame Relay Data Link Connection Identifier (DLCI) at each ingress. This allows the interior switches to use normal FR label switching – at each interior switch, the ingress port and DLCI are switched to the proper egress port and DLCI.

IP Navigator over ATM is somewhat more complicated because the packets are actually contained in multiple cells, which have to be reassembled at the egress switch. Because AAL5 is used to segment and reassemble the cells, cells from different packets cannot be interleaved on the same ATM connection. To allow the packets to be reassembled, each egress switch is assigned a unique VPI and each ingress switch sending to that VPI is assigned a unique VCI. In the interior switches, only VP switching is used; the only use for the VCI field is to identify uniquely the source switch in order to reassemble the packets at the egress switch. This technique is called VP merging, and it again allows the number of VPs in the network to scale on the order of the number of switches.

© 2000 by CRC Press LLC

Of course, this example shows only the MPT rooted at S4. Each of the other switches also has an MPT rooted at it, with leaves at every other switch. IP routing is used to determine the path each MPT uses through the network.

Once the MPTs have been established, IP Navigator uses the MPTs to route packets through the network, rather than IP routing. This allows a network operator to override routing when necessary to perform traffic engineering – to cause the packets to follow a particular route for reasons of policy. For example, in Figure 8.16, packets from S1 traveled through S2 to reach S4. However, the network administration could determine that the trunk from S2 to S4 is overutilized, and the trunks from S2 to S3 and S3 to S4 are underutilized. IP Navigator has facilities to allow the network administration to override IP routing, in this case to route the MPT branch from S1 to S3 rather than S1 to S2.

IP Navigator can also use MPTs to provide enhanced IP functionality, such as QoS support. By allocating multiple MPTs to a particular destination, with particular ATM or Frame Relay-layer QoS attributes associated with each MPT, the same QoS can then be imparted to the packets. So, best-effort packets can be sent along a best-effort MPT, while packets requiring a better QoS (such as for voice or video over IP) can be sent over a different MPT that has the required QoS at the switched layer.

IP Navigator uses OSPF as its IP routing protocol. OSPF is a link-state routing protocol which means that every switch knows the complete physical topology of the network. This makes it very easy for each MPT's root switch to determine the MPT's routing through the network, assign the necessary labels at each the leaves and at each hop, and ensure that routing loops will not occur. Routing loops must be avoided at all costs, because ATM cells and Frame Relay frames do not have any mechanism to timeout cells or frames if loops occur; they will travel through the loop for as long as it exists. Especially at higher speeds, this can cause looping cells or frames to consume considerable amounts of trunk and switch bandwidth. IP Navigator's use of OSPF makes stable loops impossible to form.

IP Navigator also includes IP multicast support by using normal ATM or Frame Relay point-to-multipoint connections to carry the multicast traffic. All required multicast replication takes places at the link layer in the switches.

As of this writing (July 1998), IP Navigator is running operationally in a number of production public IP networks. In addition, Ascend Communications is actively participating in the MPLS standardization effort, and has submitted much of IP Navigator's technology, especially MPTs, to the IETF for use in MPLS.

8.4.3 ARIS

IBM's Aggregate Route-Based IP Switching (ARIS)[19] is another control-driven method of integrating IP and ATM. ARIS took many of the concepts introduced in IP Navigator and generalized them for wider use and to remove the dependence on OSPF as the IP routing algorithm. Instead of piggybacking the MPT setup information into the OSPF routing updates, as in IP Navigator, ARIS has a separate setup algorithm that begins at the egress node of each MPT (for example, S4 in Figure 8.16) and spreads out, hop by hop, following IP routing backwards until it

© 2000 by CRC Press LLC

reaches the MPT's leaf nodes. It still depends on IP routing to set up the MPTs, but it is independent of any particular routing protocol. Because it cannot count on a link-state routing protocol, such as OSPF, to prevent routing loops, ARIS includes a separate loop prevention algorithm as a part of its MPT setup protocol. For use over ATM, ARIS also generalizes the VP-merging capability of IP Navigator's MPTs to allow VC merging as well, when used with ATM switches that include VC merge hardware. VC merging is very similar to the operation of IP Navigator over Frame Relay, in that only a single ATM VPI/VCI label is required to identify a MPT. However, this requires hardware in intermediate switches that can buffer the cells from a particular packet until they have all arrived at a switch, and then send them as an integral group of cells to the next switch. It should be noted that VC switching obviates all QoS support from the ATM network, thus should be used only for best-effort IP traffic.

As of this writing, IBM has canceled its ARIS project but has submitted ARIS' technology to the IETF for inclusion in MPLS, which it will then implement in its products.

8.4.4 Tag Switching

Tag Switching,[20] from Cisco Systems is another control-driven label switching protocol (they use the term *tag* to represent labels used for switching). Tag switching further generalizes the concepts in IP Navigator and ARIS. New features in tag switching include the following:

- Routers as well as switches take part in tag switching, which was a natural step given that routers are Cisco's primary product. Tag switching in routers allows them to participate with switches in the formation of switched paths, to simplify their own operation as packets travel though them, and, for the first time, to support real traffic engineering.
- Cisco used TCP to carry its Tag Distribution Protocol (TDP), which is necessary since Cisco supports a wide range of IP routing protocols and wanted tag switching to work with all of them. Using TCP simplified TDP's operation because it could assume TCP's reliable transport service. In addition, it allowed TDP peers to be physically separate from each other.
- They added support for routing hierarchy by introducing the concept of a *tag stack*, which allows labels to be stacked in packets. The top-most label is always the one used for tag switching through a router or switch, but they have lower-level tags for routing them through the various Internet routing domains, and within each domain, the top-most tag routes the packets through the local routers and switches. When packets enter a routing domain, a new tag is pushed on the stack, and when they leave a routing domain the top-most label is popped from the stack.
- They added the concept of the *forwarding equivalency class* (FEC), which is a set of packets that follow the same path to a particular destination.

© 2000 by CRC Press LLC

Normal IP routing has one equivalency class, which is the best match of its IP address to a network identifier in a router's routing tables. They identified a number of additional FECs, from the egress point from a network used by a number of IP routes to the set of packets belonging to a particular application at one particular destination host.

- In addition to using TDP to distribute tags, they also allow tag distribution to be piggybacked onto other protocols, such as the Resource ReSerVation Protocol (RSVP)[21] for flows that require a particular QoS, or onto the Border Gateway Protocol (BGP)[22] to support interdomain routing.

One particular criticism of tag switching is that it does not include a loop prevention algorithm when setting up tags. Cisco felt that loops, when they do form, are of short enough duration that looping packets would not consume excessive network resources. This may be true for routers, but loops lasting even several seconds at ATM speeds could be disastrous.

As of this writing, Cisco has deployed tag switching in its routers and is testing tag switching in its ATM and Frame Relay switches. Cisco has also contributed heavily to the MPLS effort in the IETF.

8.4.5 Multiprotocol Label Switching (MPLS)

The IETF's MPLS working group began its work in April 1997. Their primary goal is to produce an interoperable, multivendor approach to using label swapping to integrate IP with ATM and Frame Relay, to speed up switching in routers, and to provide unified traffic engineering functionality in both switched and routed networks. The primary technology inputs to the process are, as mentioned above, IP Navigator, ARIS, and tag switching. To quote from the working group's charter, "The working group is responsible for standardizing a base technology for using label swapping forwarding paradigm (label switching) in conjunction with network layer routing and for the implementation of that technology over various link level technologies, which may include Packet-over-SONET, Frame Relay, ATM, Ethernet (all forms, such as Gigabit Ethernet, etc.), Token Ring, and so on. This includes procedures and protocols for the distribution of labels between routers, encapsulations, multicast considerations, use of labels to support higher layer resource reservation and QoS mechanisms, and definition of host behaviors."[23]

As of this writing, all of the work is in the draft stage, which means that any and all of the details given here are subject to change before the documents are published as RFCs. However, the group has reached consensus on the broad outline of the solution.

First, it has produced draft framework[24] and architecture[25] documents. The framework document discusses MPLS technical issues and requirements and provides a broad survey of the different approaches and solutions that have been considered by the working group, including the three major inputs to the group described above in Sections 8.4.2, 8.4.3, and 8.4.4. In contrast, the architecture document describes the technical approaches that have been agreed upon to this point in the working group; it winnows through the range of possible solutions and

© 2000 by CRC Press LLC

mechanisms presented in the framework document to make specific choices. The highlights of the architectural agreements to date are provided below:

- The working group has agreed upon the basic terminology to be used, which is actually a very important step. The *label* is the basic information that is used to switch an incoming packet to an outgoing interface, possibly with a new label. The label is the same as the tag in tag switching, and can be an ATM VPI/VCI, Frame Relay DLCI, or a field in an MPLS header on point-to-point lines or LANs. Routers and switches that participate in MPLS are *Label-Switched Routers* (LSRs). They set up *Label-Switched Paths* (LSPs) to carry the IP data packets (an IP Navigator MPT is an example of an LSP). LSRs cooperate to set up LSPs by using a *Label Distribution Protocol* (LDP). LSPs are unidirectional, and, as a result, LSRs are either *upstream* or *downstream* from their neighbor, depending on the direction of traffic flow on an LSP.
- MPLS will be control-driven but will have options for being flow-driven, as well, for use in enterprise networks. MPLS' use in the Internet backbone will always be control-driven.
- As in IP Navigator, LSPs use merging in order to scale for use over wide-area Internet backbone networks.
- As in ARIS, loop prevention mechanisms are used to prevent loops from forming in LSPs.
- As in tag switching, labels can be stacked for hierarchical routing and can represent different forwarding equivalence classes.

Figure 8.17 shows an example LSP to R5. Note that the major difference between this and Figure 8.16 is that the routers are now also participating in the LSP, which saves an IP routing hop when the packets enter the FR or ATM network. In Figure 8.17, switch S2 is downstream from S1 and upstream to S4.

The working group has also produced a first draft of the Label Distribution Protocol specification,[26] which is largely a combination of ARIS' setup algorithm and tag switching's TDP. Like TDP, it also runs over TCP.

The working group is considering several quite different proposals to provide advanced traffic engineering and bandwidth management services. One provides the functionality as a part of LDP itself. The other, which proposes modifying the RSVP to carry labels, similar to tag switching's use of RSVP, is somewhat controversial because of doubts about the suitability of RVSP in WANs.

A number of additional technical issues remain to be settled before the MPLS work in the IETF can be considered complete enough for vendors to implement and attempt interoperability testing. These include encapsulation of IP packets over ATM and Frame Relay networks, the ATM classes of service for best-effort and QoS LSPs crossing ATM networks, and how to provide virtual private network functionality in MPLS, among others. However, it does promise the potential to improve greatly, in an interoperable manner, the operation of IP both across the Internet backbone and in enterprise networks.

© 2000 by CRC Press LLC

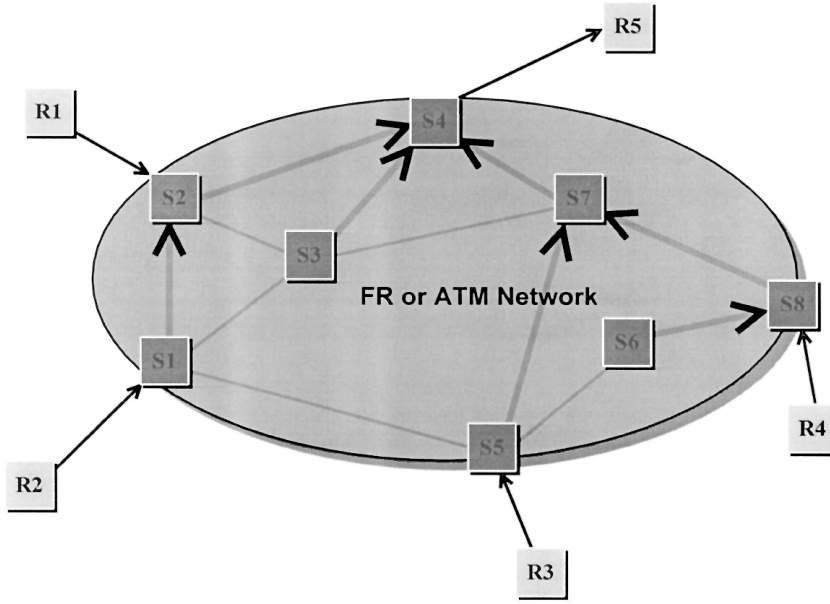

Figure 8.17 MPLS LSP to R5

REFERENCES

1. Heinänen, J., RFC 1483, Multiprotocol Encapsulation over ATM Adaptation Layer 5, July 1993, www.ietf.org.
2. ITU-T Recommendation I.363.5, B-ISDN ATM Adaptation Layer specification: Type 5 AAL, August 1996, www.itu.ch.
3. ISO/IEC Technical Report 9577 (also published as ITU-T Recommendation X.263), Protocol Identification in the Network Layer, November 1995, www.itu.ch.
4. International Standard, Information Processing Systems – Local Area Networks - Logical Link Control, ISO 8802-2: 1989 (E), IEEE Standard 802.2-1989, December 1989, www.ieee.org.
5. Postel, J., Reynolds, J., RFC 1700, Assigned Numbers, October 1994, www.ietf.org.
6. Perez, M., et al., RFC 1755, ATM Signaling Support for IP over ATM, February 1995, www.ietf.org.
7. Laubach, M., Halpern, J., RFC 2225, Classical IP and ARP over ATM, April 1998, www.ietf.org.
8. Maher, M., RFC 2331, ATM Signalling Support for IP over ATM — UNI Signalling 4.0 Update, April 1998, www.ietf.org.
9. ATM Forum af-uni-0010.001, ATM User-Network Interface Specification V3.0, September 1993, www.atmforum.com.
10. ATM Forum af-uni-0010.002, ATM User-Network Interface Specification V3.1, 1994, www.atmforum.com.

© 2000 by CRC Press LLC

11. ATM Forum af-sig-0061.000, UNI Signaling 4.0, July 1996, www.atmforum.com.
12. Luciani, J., et al., RFC 2332, NBMA Next Hop Resolution Protocol (NHRP), April 1998, www.ietf.org.
13. ATM Forum af-mpoa-0087.000, Multi-Protocol Over ATM Specification v1.0, July 1997, www.atmforum.com.
14. ATM Forum af-lane-0021.000, LAN Emulation over ATM 1.0, January 1995, www.atmforum.com.
15. Thompson, K., et al., Wide-Area Internet Traffic Patterns and Characteristics, *IEEE Network*, November/December 1997.
16. Newman, P., et al., RFC 1953, Ipsilon Flow Management Protocol Specification for IPv4 Version 1.0, May 1996, www.ietf.org.
17. H. Ahmed, et al., IP Switching for Scalable IP Services, *Proc. IEEE*, December 1997.
18. Moy, J., RFC 2328, OSPF Version 2, April 1998, www.ietf.org.
19. Viswanathan, A., et al., ARIS: Aggregate Route-Based IP Switching, Work in progress, March 1997, www.ietf.org.
20. Rekhter, Y., et al., RFC 2105, Cisco Systems' Tag Switching Architecture Overview, February 1997, www.ietf.org.
21. Braden, R., et al., RFC 2205, Resource ReSerVation Protocol (RSVP) — Version 1 Functional Specification, September 1997, www.ietf.org.
22. Rekhter, Y., et al., RFC 1771, Border Gateway Protocol 4 (BGP-4), March 1995, www.ietf.org.
23. IETF MPLS Working Group charter, http://www.ietf.org/html.charters/mpls-charter.html.
24. Callon, R., et al., A Framework for Multiprotocol Label Switching, Work in progress, November 1997, www.ietf.org.
25. Rosen, E., et al., Multiprotocol Label Switching Architecture, Work in progress, March 1998, www.ietf.org.
26. Andersson, L., et al., Label Distribution Protocol, Work in progress, March 1998, www.ietf.org.

© 2000 by CRC Press LLC