Internetworking with ATM WANs

John David Cavanaugh
Timothy J. Salo
Minnesota Supercomputer Center, Inc.*

December 14, 1992

Abstract

High-speed (155–622 Mb/s) wide-area networks based on the Broadband Integrated Services Digital Network's (B-ISDN) Asynchronous Transfer Mode (ATM) technology are currently being tested, and commercial service offerings are expected soon. This paper is an overview of ATM, ATM adaptation layers, and the Connectionless Broadband Data Service (CBDS). It discusses internetworking strategies and issues that arise when using ATM networks to carry the TCP/IP protocol suite.

1 Introduction

The Internet Protocol (IP) family (IP, TCP, UDP, Telnet, SMTP, FTP, etc.) has met with great success during the last decade. It is used on many different types of computers, from PCs to supercomputers. It is used in networks that fit in a single room, networks that serve a campus or corporation, and networks that span the globe. It uses a variety of media—including Ethernet[1], Token Ring, FDDI, X.25, high-speed computer channels, switched and leased telephone lines—ranging in speed from a few hundred bits per second to hundreds of megabits per second. It is used by a wide range of applications—file transfer, electronic mail, file sharing, and graphical user interfaces, to name just a few. It has proven effective in all these varied circumstances.

Work is now under way to adapt the IP protocol family to emerging network technologies that offer very-high-speed (up to the gigabit per second range) data transfer over long distances. Two of these new technologies are Asynchronous Transfer Mode (ATM) and Switched Multi-Megabit Data Service (SMDS). Both are public, switched data services and are expected to develop into networks of international scope, similar to the telephone network. ATM is connection-oriented; SMDS is connectionless. ATM is expected to be available at 155 and 622 Mb/s in the public network; SMDS is currently being tested at 1.5 and 45 Mb/s and is expected to reach 150 Mb/s.

These new technologies pose some challenges to the IP protocol family, mainly to the IP and TCP layers. A number of issues must be resolved before the IP protocol family can be used widely over these new networks. Some of these issues arise simply because the network media are new; some are due to the structure of the networks; some are caused by the speed of the networks. These issues (discussed in detail in sections 3 and 4) include the following:

- appropriate internetworking strategies
- format of encapsulation of IP datagrams
- network definition
- management of ATM connections
- TCP performance issues
- efficiency of datagram transmission

*The research described herein was sponsored by DARPA through the U. S. Army Research Office, Department of the Army, contract number DAAL03-91-C-0049. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by either DARPA or the U. S. Army.

[1] Ethernet is a registered trademark of Xerox Corporation.

Although many of these issues are common to several network technologies, this paper focuses on using the IP protocol family on ATM networks.

1.1 Organization of This Paper

Section 2 is an overview of ATM. Section 2.2 discusses ATM adaptation layers. Section 2.3 discusses the Connectionless Broadband Data Service (CBDS); readers who are already familiar with these topics may wish to skip these sections. Internetworking strategies for IP over ATM are discussed in section 3. Networking issues related to IP and ATM are discussed in section 4. Section 5 is a summary of the discussion and presents recommendations and conclusions.

Appendix A is a list of the acronyms used in this paper. Appendix B lists the CCITT recommendations for Broadband ISDN (B-ISDN).

2 An Overview of ATM

Asynchronous Transfer Mode[2] (ATM) is a technique for multiplexing fixed-length cells from a variety of sources to a variety of remote locations. ATM is capable of moving data at a wide range of speeds, but its usual application will be to carry data over optical fiber at very high speed (100–1000 Mb/s). It is capable of handling data from a variety of media (e.g. voice, video, and data) using a single interface and can multiplex a number of connections onto that interface.

ATM is a connection-oriented protocol. Connections are established between ATM service users, whether they are attached to the ATM network (e.g. a workstation with an ATM interface) or part of it (e.g. a call setup process running on an ATM switch). Connections can be switched, semi-permanent, or permanent. *Signalling procedures* are used to set up switched calls. These signalling procedures correspond to the dialing of a telephone. Semi-permanent and permanent connections are established through administrative procedures.

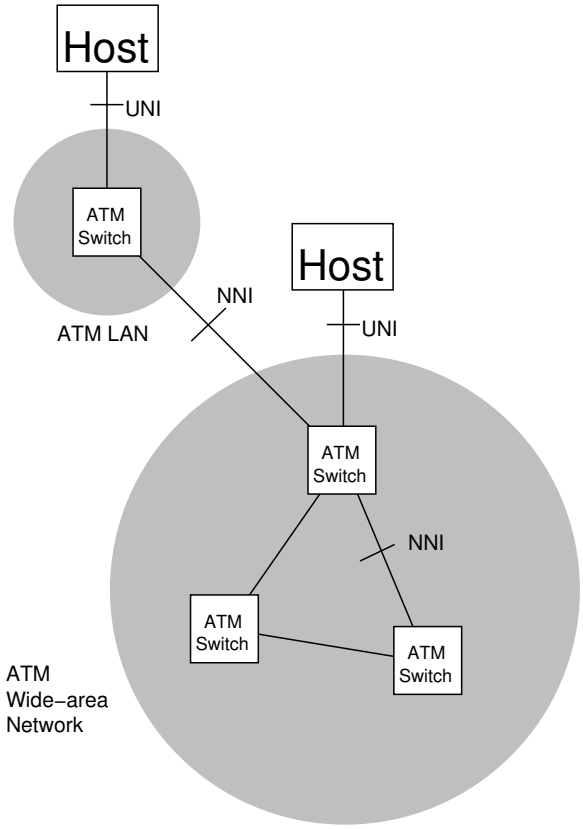

Figure 1: ATM Network Structure

The ATM service in wide-area networks is expected to run over a SONET carrier at SONET speeds (other media may be used in LANs). The speeds that are important to this discussion are OC-3c[3] (155.520 Mb/s), and OC-12c (622.080 Mb/s).

Each ATM connection operates at a certain *grade of service* (GOS). A particular GOS is characterized by such parameters as cell loss and cell delay variation. The GOS is negotiated between the user and the network when the connection is established. ATM operates on a

[2] Asynchronous Transfer Mode can be used to mean either the general technique of transferring data via fixed length cells or the particular version of that technique specified by the CCITT as the transport protocol for B-ISDN. In this paper, we use it in the latter sense.

[3] SONET's transmission rates are defined as multiples of 51.840 Mb/s OC-1 channels. An OC-3 channel is capable of carrying 3 OC-1 channels, or 3×51.840 Mb/s, 155.520 Mb/s. The designation 'c' indicates that the channels are concatenated; that is, they operate as a single channel rather than as n multiplexed independent channels. An OC-3 channel comprises three separate OC-1 channels, while an OC-3c channel is a single channel with three times the capacity of an OC-1 channel.

"best effort" basis; cells with errors, or that encounter congestion, are silently dropped. It is the responsibility of higher protocol layers to notice that cells are missing and, if required, to arrange for retransmission of the lost cells. A sequence of cells in an ATM connection will be received in the same order as it was transmitted; ATM guarantees that cells will not be misordered.

There are two types of connections: point-to-point and multipoint. Point-to-point connections are like ordinary two-party telephone calls; they connect two ATM service users. Multipoint connections are like conference calls; they connect more than two ATM service users. Multipoint connections can be used to provide a multicast facility.

Figure 1 shows a simplified ATM network. Network services are provided by ATM switches within the network. In an ATM LAN, the ATM switch will act as the network hub. Hosts are attached to the ATM network at the *user–network interface* (UNI). ATM switches are interconnected at the *network node interface* (NNI). Two distinct NNIs may be needed: one for the interconnection of ATM switches within the public network and a second for the attachment of ATM LAN switches to the public network.

2.1 ATM Cell

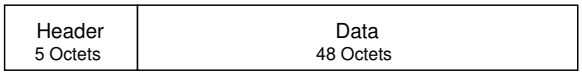

Figure 2: ATM Cell Structure

ATM transfers data in fixed-length units called *cells*. An ATM cell is 53 octets long—5 octets are header information and the remaining 48 are data. If the payload in a cell does not require the full 48 octets, padding is used (by the layer above ATM) to fill the cell. The 48-octet size represents a compromise between the demand of voice traffic for quick access to the network and the demand of data traffic for large data units. Figure 2 shows the layout of the ATM cell; Figure 3 shows the layout of the ATM header.

The fields in the header have the following uses:

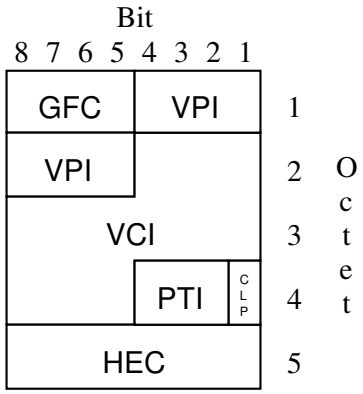

Figure 3: ATM Header

GFC Generic Flow Control. 4 bits. Used by the flow control mechanism at the UNI. Mechanisms are yet to be determined.

VPI Virtual Path Identifier. 8 bits. Used for directing cells within the ATM network (see discussion below).

VCI Virtual Channel Identifier. 16 bits. Used for directing cells within the ATM network (see discussion below).

PTI Payload Type Indicator. 3 bits. Identifies the type of data being carried by the cell. See Table 1 for values.

CLP Cell Loss Priority. 1 bit. If this bit is set (equal to 1), the cell has low priority and is subject to being discarded when the network is under stress. If it is not set (equal to 0) the cell has higher priority and is less likely to be discarded.

HEC Header Error Correction. 8 bits. Generated and inserted by the physical layer. Serves as a checksum for the first 4 octets of the ATM header. It can correct single-bit errors and detect some multiple-bit errors.

ATM cells flow along entities known as *virtual channels* (VCs). A VC is identified by its *virtual channel identifier* (VCI). All cells in a given VC follow the same route across the network and are delivered in the order they were transmitted. A VC between two users can carry data and signalling information. A VC between a user and the ATM network can be used for signalling or for administrative purposes. Three VCs—numbers 0, 1, and 2—are reserved for

special purposes. 0 is for unassigned cells; 1 indicates the meta-signalling VC; 2 indicates the general broadcast signalling VC.

VCs are transported within *virtual paths* (VPs). A VP is identified by its *virtual path identifier* (VPI). VPs are used for aggregating VCs together or for providing an unstructured data pipe. Depending on the requirements of the user and the network, it is possible that not all bits of the VPI and VCI will be significant. Bits that are not significant are set to zero.

The VPI and VCI, taken together, form a 24-bit *protocol connection identifier* (PCI). The PCI identifies a particular call or connection and is used for routing cells across the network and demultiplexing cells at the destination. Cells with PCI equal to zero are *unassigned* or unused. They carry no user data.

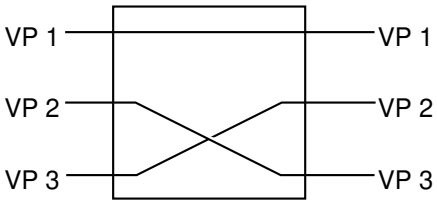

Figure 4: VP Switch

VPs and VCs are subject to switching within the ATM network. A VP switch can redirect a VP, perhaps reassigning the VPI, but keeps the VCs within the VP intact (see Figure 4).

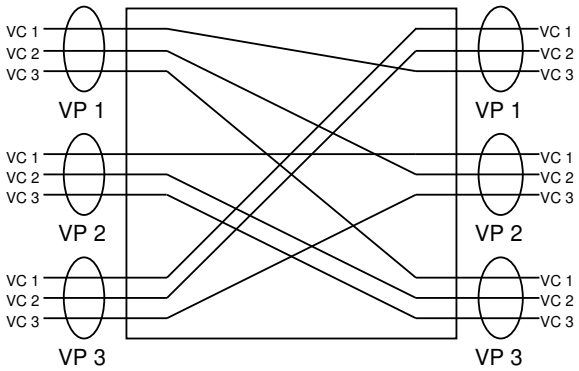

Figure 5: VC Switch

A VC switch must terminate VPs and can switch the VCs within a VP independently of each other (including reassigning their VCIs; see Figure 5).

All this switching means that the VPI and VCI at different ends of a connection may not be the same. That is, each end point of a connection (and there may be more than two for a multipoint connection) may refer to the same connection with a different VPI or VCI.

The type of information carried in each cell is denoted by the PTI field. The meanings of the different values of this field are shown in Table 1.

000	User data, congestion not experienced, Segmentation Data Unit (SDU) type 0
001	User data, congestion not experienced, SDU type 1
010	User data, congestion experienced, SDU type 0
011	User data, congestion experienced, SDU type 1
100	Segment Operation and Maintenance (OAM) F5 flow-related cell
101	End-to-end OAM F5 flow-related cell
110	Resource management cell
111	Reserved for future use

Table 1: Payload Type Indicator Encoding

Note that in user data cells (PTI = 0XX), the value of the third bit indicates the Service Data Unit (SDU) type. This is used by AAL 5 to indicate the end of a data unit (see section 2.2.2).

The use of cells with PTI types 100, 101, 110, and 111 is for further study. Cells with PTI types 100 and 101 are Operation and Maintenance (OAM) F5 cells; these are cells that carry administrative messages related to the operation and maintenance of the virtual circuit that carries them.

2.2 ATM Adaptation Layers

In order to carry data units longer than 48 octets in ATM cells, an adaptation layer is needed. The *ATM adaptation layer* (AAL) provides for segmentation and reassembly of higher-layer data units and for detection of errors in transmission.

Since the ATM layer simply carries cells without concern for their contents, a number of different AALs can be used across a single ATM interface. The end points of each connection must agree on which AAL they will use, but the network need not be concerned with this.

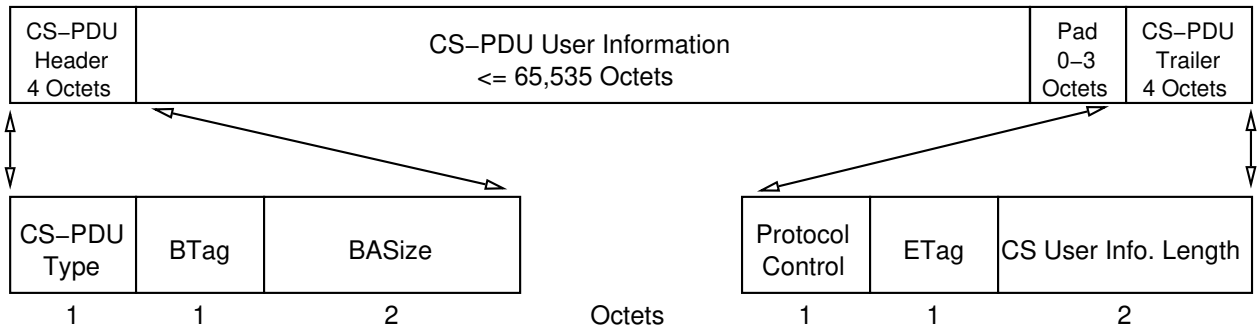

Figure 6: AAL 3/4 Convergence Sublayer (CS) PDU Structure

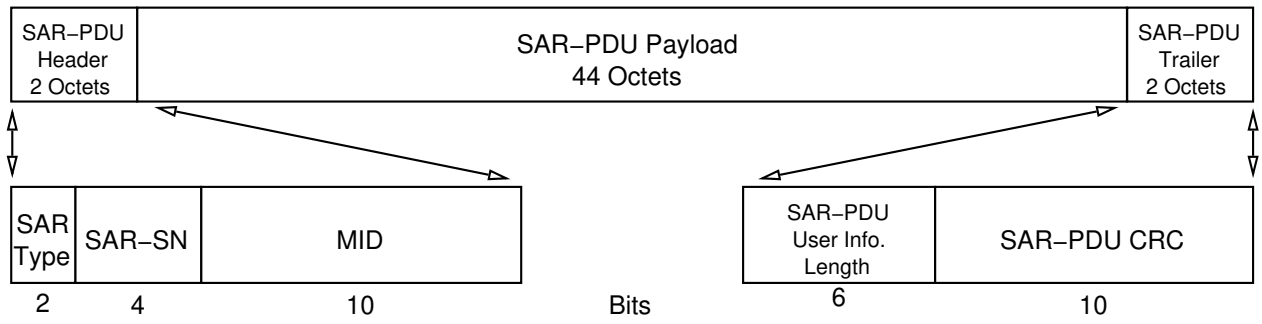

Figure 7: AAL 3/4 Segmentation and Reassembly (SAR) PDU Structure

Five AALs have been accepted for consideration by the CCITT. They are referred to as AAL 1–5. AAL 1 is meant for constant-bit-rate services (e.g. voice). AAL 2 is meant for variable-bit-rate services with a required timing relationship between source and destination (e.g. audio or video). AAL 3 was originally meant for connection-oriented variable-bit-rate services without a required timing relationship; it has now been merged with AAL 4. AAL 3/4 and 5 are meant for connectionless services (e.g. connectionless data). Only AALs 3/4 and 5 are of interest for IP networking.

2.2.1 AAL 3/4

AAL 3/4, the *variable bit rate* (VBR) adaptation layer, defined in CCITT recommendation I.363, is defined for services (e.g. data) that require "bursty" bandwidth. AAL 3/4 actually comprises two sublayers, the *convergence sublayer* (CS) and the *segmentation and reassembly* (SAR) sublayer. The CS protocol data unit (PDU) structure is shown in Figure 6; the SAR PDU structure is shown in Figure 7.

The pad field in the CS-PDU ensures that the length of the PDU is divisible by 4. The value of the pad field is binary zeroes.

The fields in the CS header and trailer have the following uses:

CS-PDU Type 1 octet. Use under study. When the upper layer is connectionless, this field is set to zero.

BTag 1 octet. Use under study. Meant to be used in conjunction with ETag.

BASize 2 octets. Contains the length in octets of the CS-PDU user information.

Protocol Control 1 octet. Use under study. When the upper layer is connectionless, this field is set to zero.

ETag 1 octet. Use under study. Meant to be used in conjunction with BTag.

CS-PDU User Information Length 2 octets. The length of the user information in the CS-PDU (not including pad).

As indicated by the number of fields described as "Use under study," the formulation of the AAL 3/4 CS is still incomplete.

The fields in the SAR header and trailer have the following uses:

SAR Type 2 bits. Differentiates the parts of a segmented message. Values are:

01	beginning of message (BOM)
00	continuation of message (COM)
10	end of message (EOM)
11	single-segment message (SSM)

SAR-SN 4 bits. Indicates the position of the segment within the message[4]

MID 10 bits. Identifies the message. All segments of a given message will have the same MID value.

SAR-PDU User Information Length 6 bits. Gives the length of the user information in the SAR-PDU.

SAR-PDU CRC 10 bits. A CRC that covers the entire SAR-PDU.

When a network node has a user datagram to transmit, it first converts it to a CS-PDU by adding the CS-PDU header, pad (if necessary), and trailer. It then splits the CS-PDU into 44-octet chunks and converts each chunk to a SAR-PDU by adding the SAR header and trailer, and transmits them, one SAR-PDU per ATM cell. If necessary, the last chunk of the CS-PDU is padded to fill out the last SAR-PDU.

When SAR-PDUs are received, they are reassembled into CS-PDUs, with the receiver using the SAR Type, SAR-SN, and MID fields to ensure that the chunks return to their proper places. The receiver verifies that the reassembled CS-PDU is intact using the fields in the CS-PDU header and trailer (e.g. the length of the CS-PDU as reassembled by the SAR sublayer is compared to the length value in the CS-PDU trailer).

[4] The SAR-SN indicates the sequential position, modulo 16, of the segment in the message.

2.2.2 AAL 5

AAL 5, the Simple and Efficient Adaptation Layer (SEAL) [1], attempts to reduce the complexity and overhead of AAL 3/4. It eliminates most of the protocol overhead of AAL 3/4. Like AAL 3/4, AAL 5 comprises a convergence sublayer and a SAR sublayer, although the AAL 5 SAR sublayer is essentially null. The structure of the AAL 5 CS PDU is shown in Figure 8.

The pad field in the CS-PDU ensures that the total length of the CS-PDU (including pad and trailer) is divisible by 48 and that the CS-PDU trailer resides in the last eight octets of the last 48-octet chunk. The value of the pad field is binary zeroes.

The fields in the AAL 5 CS-PDU trailer have the following uses:

Control 16 bits. Use under study.

Length 16 bits. The length of the user information in the CS-PDU (not including pad).

CRC 32 bits. A 32-bit CRC (CRC-32; the same as that used in FDDI and Fibre Channel) covering the data and pad.

When a network node has a user datagram to transmit, it first converts it to a CS-PDU by adding the pad (if necessary) and trailer. Then it breaks the CS-PDU into 48-octet SAR-PDUs and transmits each in an ATM cell on the same virtual channel. The last SAR-PDU is marked so that the receiver can recognize it. Since there is no AAL 5 SAR header, an end-of-frame indication in the ATM cell header is required. The proposed end-of-frame indication is an SDU type of 1 (binary value '0X1') in the Payload Type Indicator (PTI) field.

The receiver simply concatenates cells as they are received, watching for the end-of-frame indication. When it is seen, the receiver checks the length and CRC and passes the PDU up to the next higher layer. The higher layer is responsible for ignoring PDUs with CRC errors. Some applications may discard PDUs with errors; others (e.g. voice or video, which can

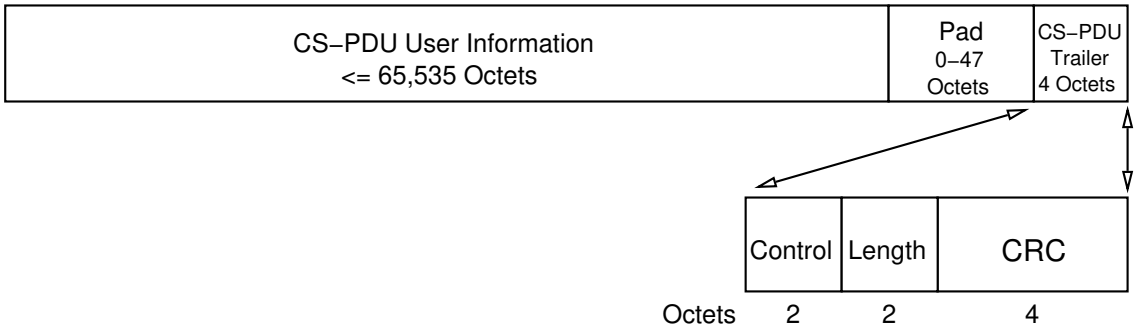

Figure 8: AAL 5 PDU Structure

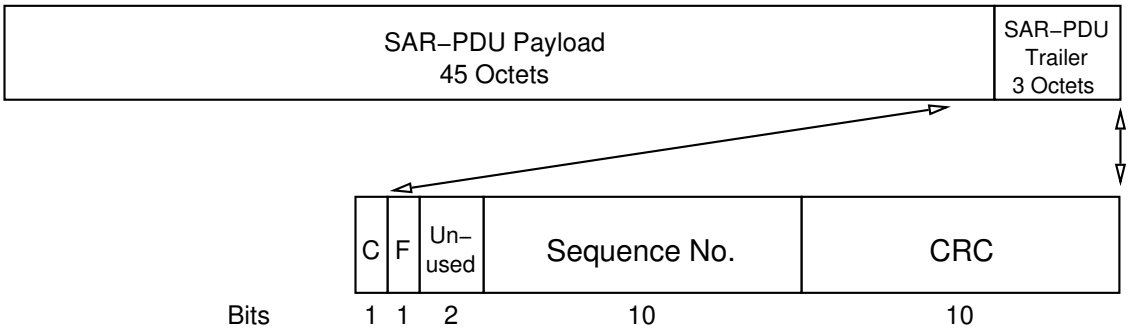

Figure 9: BBN SAR PDU Structure

tolerate some errors) may choose to use them, or to do error correction.

2.2.3 Other AALs

In addition to AAL 1–5, two other AALs have been defined. The first, BLINKBLT [13], provides the complete OSI data link layer service across an ATM network. Unlike the other AALs, it contains provisions for recovery of lost cells.

BLINKBLT comprises two sublayers, the cell and frame sublayers, which correspond roughly to the CS and SAR of AAL 3/4. Eight PDU types are defined; two are for the transport of user data, the others are control messages.

BLINKBLT is unique in that, unlike the other AALs, it is connection-oriented. Connections are established by a three-way handshake and closed by a similar mechanism. The connection is considered to consist of two one-way circuits, each with its own sequence number space. Within each half-circuit, the receiver controls the data rate by granting credits to the transmitter. Each correctly-received cell is explicitly acknowledged; cells that are lost or received with errors are retransmitted. Cells are protected by a 10-bit CRC[5] and frames are further protected by a 32-bit CRC.

The second alternative AAL, the BBN SAR protocol [11], is an experimental segmentation-and-reassembly protocol that was defined at Bolt Beranek and Newman, Inc. (BBN). The BBN SAR was designed to replace the existing profusion of AALs with a single protocol. It was designed to work with an undefined (and possibly empty) convergence sublayer. These are the four main features of the BBN SAR:

- detection of lost or misordered cells via a sequence number
- error correction via a 10-bit CRC
- demarcation of higher-layer data units via a frame bit

[5] The same CRC is used by AAL 3/4.

- provision for insertion of control cells into the data stream

Control information for the BBN SAR is held in a three-octet trailer inserted into each ATM cell. Figure 9 shows the layout of the trailer and its position in the cell.

The fields in the BBN SAR trailer have the following uses:

C 1 bit. The control-cell bit. Set to 1 if the cell is for control; set to 0 if the cell carries user data.

F 1 bit. The frame bit. Set to 1 if this is the last cell of a frame; set to 0 otherwise.

Unused 2 bits. Currently unused.

Sequence No. 10 bits. A counter that is incremented (modulo 1024) by 1 for each cell sent over an ATM connection.

CRC 10 bits. A CRC[6] for detection and correction of errors[7].

Transmission and reception of SAR-PDUs is similar to that of AAL 3/4 and 5. The transmitter breaks the higher-layer data unit into chunks, appends the SAR trailer, and transmits them, while the receiver reassembles the chunks into the original data unit. The main difference is the introduction of a window based on the sequence number.

The trailing edge of the window is the next sequence number the receiver expects. The leading edge is set by the application to the trailing edge plus the number of cells the receiver is prepared to buffer. During normal data transfer, the window moves forward by one each time a cell is received. When a cell is lost or misordered, the window freezes. The receiver accepts and stores cells whose sequence numbers fall within the window until the window fills, a time-out occurs, or the missing cell is received. In either of the first two cases, the receiver reports a lost cell to the next higher layer and repositions the window past the lost cell. In the third case, the receiver passes all the received cells to the higher layer and repositions the window.

Both BLINKBLT and the BBN SAR were submitted to ANSI for consideration. However, ANSI did not choose either as the basis for a standard, so neither was submitted to CCITT.

[6] Again, the same CRC as is used by AAL 3/4.

[7] In AAL 3/4, the correction of errors is optional, in the BBN SAR, it is required.

2.3 Connectionless Broadband Data Service

In order to provide a connectionless data service over an ATM network, the Connectionless Broadband Data Service (CBDS) has been defined in draft recommendation I.364. An example ATM internetwork supporting CBDS is shown in Figure 10.

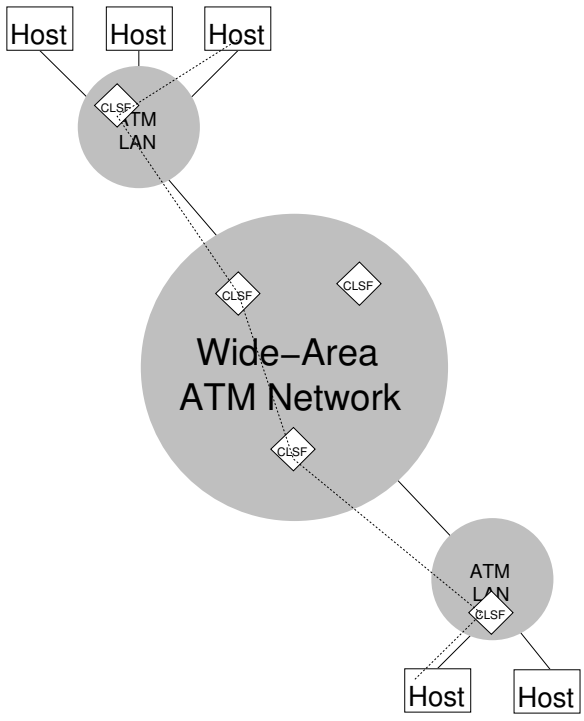

Figure 10: CBDS Operation

As shown in Figure 10, a number of *connectionless service functions* (CLSFs) exist throughout the network. The CLSFs act as cell-based routers, forwarding connectionless data units (CL-PDUs) across the network.

To send a CL-PDU, an ATM station does not attempt to make an ATM connection to the destination. Instead, it sends the PDU to a

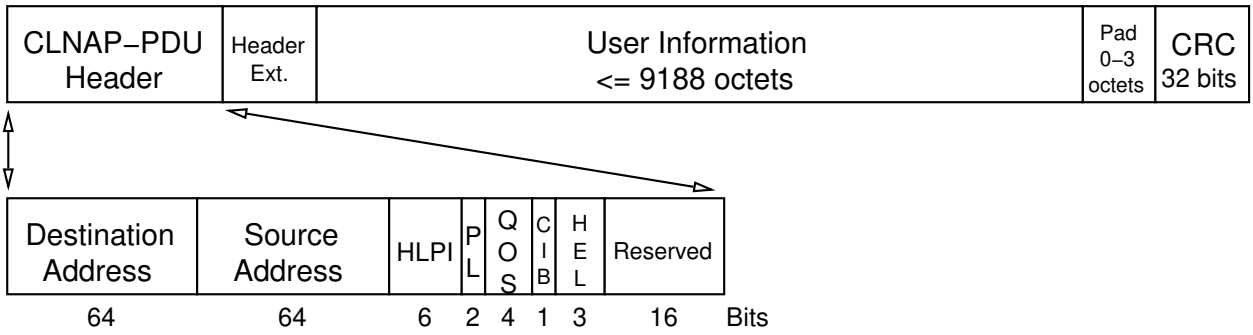

Figure 11: CLNAP PDU Structure

CLSF, opening a connection to it if necessary (it is probably preferable that the station already have an open (semi-) permanent connection to the CLSF). The receiving CLSF relays the CL-PDU to another CLSF closer to the destination, and so the CL-PDU "hops" its way across the network until it reaches a CLSF that can deliver it to the destination. The CLSFs do not reassemble the CL-PDU at each hop, but "stream" the ATM cells comprising the CL-PDU. The CL-PDU is only reassembled at its destination.

The protocol used by the CLSFs is the Connectionless Network Access Protocol (CLNAP). The structure of the CLNAP PDU is shown in Figure 11.

The fields in the CLNAP PDU have the following uses:

Destination Address 64 bits. The E.164 address of the destination of the data.

Source Address 64 bits. The E.164 address of the source of the data.

HLPI Higher Layer Protocol Identifier. 6 bits. Indicates the entity that is to receive the data at the destination. Carried transparently across the network.

PL Pad Length. 2 bits. Indicates the number of pad octets in the PDU.

QOS Quality of Service. 4 bits. Indicates the quality of service requested by the source. For further study.

CIB CRC Indicator Bit. 1 bit. Set to 1 to indicate the presence of the optional CRC. For further study.

HEL Header Extension Length. 3 bits. Indicates the number of 32-bit words in the header extension field.

Reserved 16 bits. Set to zero.

Header Extension Variable length. For further study.

User Information Variable length. The data being transported on behalf of the CLNAP user.

Pad Padding. 0–3 octets. Sufficient padding to make the length of the user information plus pad divisible by 4.

CRC Cyclic Redundancy Code. 32 bits. For further study.

The format of the source and destination addresses is as specified by E.164. E.164 is a CCITT recommendation that specifies the worldwide ISDN numbering plan.

The draft recommendation specifies that the CLNAP-PDUs would be encapsulated by AAL 3/4 to be transmitted across an ATM network.

2.4 Unresolved ATM Issues

ATM is a young and evolving standard. Significant parts are still not in place. Some of these have effects only within the ATM network;

others are more visible. This section discusses a couple of the most significant uncertainties visible to upper layers.

The most obvious unresolved issue for ATM is the question of AALs. No AAL standard is complete—the AALs that have been defined are still subject to change—and no AAL has been approved for use with IP. This is an important issue because network devices that wish to interoperate have to use the same AAL (and, of course, agree on what that means).

The setting of standards for addressing and signalling within ATM networks is also important. Until addressing and signalling are defined, ATM cannot be widely used, as equipment from different vendors will probably not interoperate. One effect of this is that the interoperation of ATM LANs with the public ATM network cannot be defined. We expect that ATM LANs will eventually appear to be a part of a large, seamless ATM network (see also section 3.1), but in the absence of standards for signalling procedures, they can only be connected to the public ATM network through network-level devices (routers).

3 Internetworking Strategies

The purpose of computer networks is to allow hosts to intercommunicate. To this end, the hosts have to be connected to the network. In the usual model, a host is connected to some sort of LAN, with LANs being connected into an *internetwork* by a wide-area network. An example of such an internetwork is shown in Figure 12.

We assume that this internetwork model applies (i.e. that hosts will continue to be connected to LANs, and that the LANs will be interconnected by wide-area networks) even with the advent of high-speed ATM networks. These are our reasons:

- A connection to the public ATM network is expected to be expensive; the internetwork model allows the expense to be shared by a large number of users.

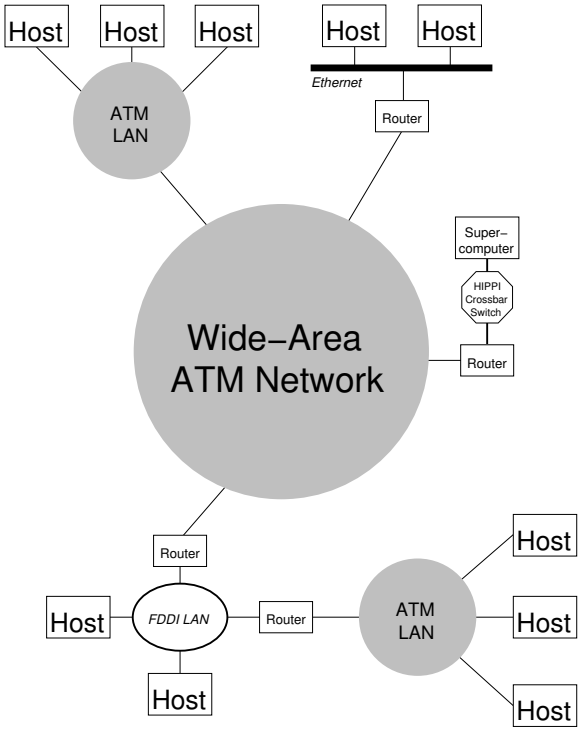

Figure 12: ATM-based Internetwork

- It is preferable for traffic between machines at a single site to cross a LAN rather than the public ATM network.
- There may be hosts for which an ATM interface is not available.
- There are many existing networks that use non-ATM media; these networks will still be in use after ATM networks are built.

3.1 ATM LAN–WAN Connection

As the public carriers are experimenting with ATM and installing ATM switches, vendors are producing small ATM switches for use in LANs. These devices are different from other LAN technologies (e.g. Ethernet, FDDI, HIPPI) in that they can potentially interoperate directly with ATM switches in the public network. An interesting question is whether, once standards for addressing, signalling, billing, etc. are in place, local ATM switches will be developed so that they can interoperate with the public network.

If they are, we can expect that ATM LANs will

be integrated more or less seamlessly into the public ATM network, so that a host on an ATM LAN will be able to open a connection directly to another host on a different ATM LAN without concerning itself with the public network that interconnects the LANs. This arrangement is similar to the use of today's PBX. Its advantage is that a host on an ATM LAN could make direct ATM connections to remote hosts across the public ATM network but still communicate with local hosts without having to cross the public network.

Its disadvantages are the complexity of the LAN switch and the assignment of addresses. The NNI of the ATM LAN switch has to conform to the standards of the public ATM network if integration is to be complete. In contrast, a non-integrated ATM LAN switch does not need a NNI at all, but would be connected to the outside world through a router attached to a UNI on the switch. Also, if ATM LAN nodes are directly addressable from the public ATM network, the public network must be cognizant of the ATM-level addresses of those nodes. This implies a high degree of cooperation between the LAN and the public network, particularly in the assignment of addresses. In one scenario, ATM addresses (assigned in accordance with CCITT recommendation E.164) are administered by the operators of the public network and anyone wishing to connect an ATM LAN to the public ATM network has to request addresses. In another [21], E.164 addresses are used in the public network, but layer 3 addresses (e.g. IP addresses) are used in LANs, with a switch at the boundary being responsible for the required translations.

If ATM switches are not fully integrated, they will, like other types of LANs, need some intermediate device (bridge or router) between them and the public ATM network. This case is discussed below.

3.2 Bridging/Routing

It is tempting to transport high-speed data from existing channels (e.g. HIPPI) by extending the channel across the ATM network. The signals and data would simply be packed into ATM cells, shipped across the network, and turned back into HIPPI signals and data by an identical device on the other side. This approach is simple and direct. However, it has drawbacks that make it impractical:

- Performance limitations because of end-to-end propagation of signals.
- The need for compatible equipment on both ends of the connection.
- Potential for loss of HIPPI signals (e.g. READY) as they cross the network.

We believe that connecting to the public ATM network through a router, rather than a bridge, has the advantage of providing isolation and increasing interoperability.

Isolation: A router provides isolation from the data link layer of the public ATM network and thereby from any low-level misbehavior on the part of the public network or an attached host or remote network. Also, the low-level (data link or physical layer) addresses of network nodes behind the router are independent of those within the public ATM network, so LAN administrators could select addresses without worrying about the public network's numbering policy. A router can also perform network-level filtering of packets, providing an important security checkpoint.

Interoperability: Given that standards for IP exist for the media involved, a router improves interoperability. Equipment from different vendors can interoperate, and there is no requirement for LANs at the ends of a connection to be of the same type. For instance, a packet could originate at a host, cross a HIPPI LAN to a router, be sent across the public ATM network to a second router (built by a different manufacturer), and cross an ATM LAN to its destination.

4 ATM/IP Networking Issues

The IP protocol suite is used on a very large number of computers of different types. As local- and wide-area ATM networks become available,

they will be expected to carry IP traffic. This section examines some of the issues involved in using ATM networks, especially wide-area ATM networks, for IP.

4.1 Format of IP Encapsulation

If IP datagrams are to be transported within ATM cells, there must be agreement on the format of the encapsulation. There is currently no standard for sending IP over ATM. A couple of proposals have been put forward [9, 15], and the Internet Engineering Task Force (IETF) has taken up the issue, but a definitive answer is probably some time away. The basic issues are the choice of AAL and a method for identifying the higher-layer protocol.

The choices for adaptation layer are AAL 3/4 and AAL 5. The first proposals specified AAL 3/4, but AAL 5 seems to be the current favorite because of its simplicity and the data integrity offered by its 32-bit CRC.

Three proposals for identifying the higher-layer protocol have been made. The first uses a different virtual circuit for each protocol. Its advantage is that the format of the transmitted data is very simple; its disadvantage is the static nature of the assignment. The second proposal includes a *network level protocol ID* (NLPID) at the start of each packet; protocols without an assigned NLPID[8] are identified by an IEEE 802.1a Subnetwork Access Protocol (SNAP) header. The third proposal identifies the protocol by using an IEEE 802.2 *logical link control* (LLC) header. Again, provision is made for use of a SNAP header for protocols that cannot be identified by the LLC header.

4.2 IP Routing

The first part of an IP address specifies a *network number*. In general, a network number corresponds to a physical network, although the concept of *subnetting* allows a single network to be split into a number of physically distinct *subnets*. Physically distinct networks are interconnected by *routers*, which forward IP datagrams between (sub-)networks. IP assumes that if two hosts have IP addresses with different (sub-)network parts, it is not possible to transmit data directly between them.

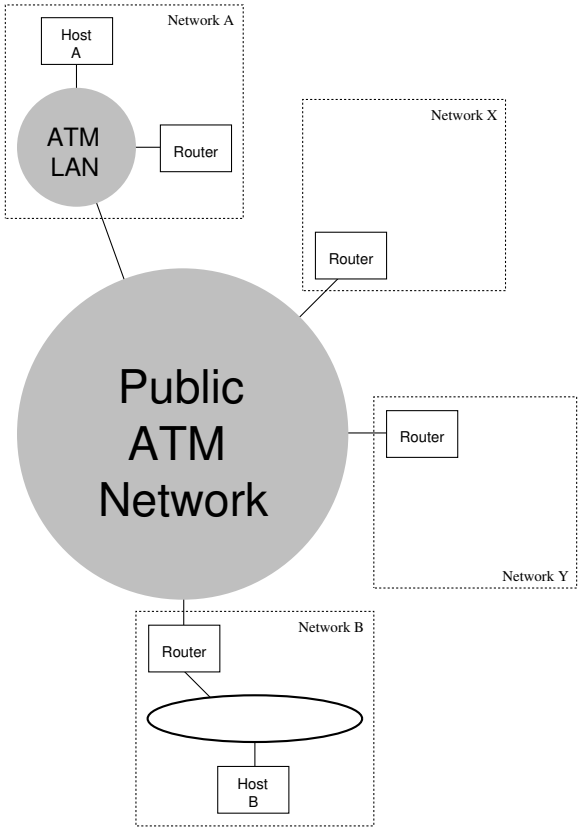

Figure 13: IP Routing Example

However, with a large number of networks all being connected to one public network, this may not be true. If both hosts have interfaces to the public network, they can, indeed, exchange data directly. As shown in Figure 13, Host A can, in fact, make an ATM connection directly to the router in network B, which means that Host B is effectively one hop away from Host A. If Host A is not aware of this, however, it will send its datagram to the router within its own network for forwarding. This router may not be aware of the router in network B and send the datagram to another router instead. In this way, the datagram may make several transits of the public ATM network on its way from Host A to Host B, when one would have been sufficient.

This problem is not unique to B-ISDN ATM networks; SMDS faces the same difficulties. In

[8] NLPID values are administered by ISO and CCITT. Values are defined in [17].

fact, a technique called *shortcut routing* [26] has been proposed for use in SMDS networks and may prove useful in ATM networks.

4.3 IP Address Resolution

Each host (or router) attached to an ATM network has one or more IP addresses. Each network interface also has an ATM address. A host that has a datagram to send knows the IP address of the destination; before it can send the datagram, it must learn the corresponding ATM address to locate (or create) a virtual channel that it can use to send the datagram.

On most LANs, a protocol known as the *address resolution protocol* (ARP) is used for this purpose. An ARP request is broadcast to all hosts on the LAN and the host that owns the address returns a response. Broadcasting is impractical (if not impossible) on a network as large as the public ATM network, so some other method of address resolution will have to be used.

Several different techniques be used. First, each network node keeps a table giving the correspondence between IP and ATM addresses for all the nodes on the network. For a network of any size, this is a huge administrative headache and clearly impractical. Second, a multipoint ATM connection is made to all nodes on a given network. This is practical if the network is small or if a group address is defined for all the nodes belonging to a particular IP network. The ARP request (and reply) is transmitted on the resulting multipoint VC. Third, a technique called *directed ARP* [12] uses the normal ARP packet formats but allows routers to forward an ARP request to another machine when there is reason to believe that it is capable of answering the request. Directed ARP allows a network node to test whether a target address is on the local physical network by simply sending an ARP request.

This difficulty with address resolution is not unique to ATM networks; SMDS faces the same problem.

4.4 ATM Connection Management

When switched connections are available, network hosts may want to make temporary connections to each other. How should these connections be managed?

One obvious solution is to "nail up" connections between hosts. These permanent connections would, like leased lines, always be available. This solution will probably be used between some hosts that have heavy traffic between them. For example, a permanent multipoint connection can be used to connect all the routers for a particular IP network that have interfaces to the public ATM network; this connection serves as a multicast server between the routers. However, it is not a general solution because it is not practical to nail up connections between a given host and all the other hosts it with which it might communicate.

Another solution is to establish connections on an "as needed" basis. When a host has a datagram to send, it checks to see whether it currently has an open connection to the destination. If it does, it uses the connection to send the datagram. If not, it stores the datagram, opens a connection, and then sends the datagram. If it cannot open a connection, it discards the datagram. If a connection is idle for a given time, it closes it. This scheme is already in use over X.25 networks. The drawback is that it increases the time a host takes to send a packet.

Ultimately, there will be a need for both types of connections. Hosts that constantly exchange a lot of data (e.g. routers) require permanent connections, while switched connections are appropriate for use between hosts that rarely communicate (e.g. only require connections for email transfer). The tariffs imposed for use of the public ATM network will also have an effect. High prices for permanent connections will encourage the use of switched connections and vice versa.

4.5 TCP Considerations

There is potential for disharmonious interactions between ATM and higher-layer protocols.

Cell loss: ATM discards cells when it encounters congestion. When an upper layer (e.g. TCP) notices that data was lost, it retransmits any unacknowledged data. The retransmitted data may be considerably more than was discarded. This has the potential for increasing the congestion, with obvious unfortunate consequences. Techniques such as slow-start and exponential retransmit timer backoff [18] should alleviate the damage, but research in live networks is needed.

TCP window management: Van Jacobson [20] states that standard TCP window algorithms run into trouble when the bandwidth $\times$ delay product of a link exceeds 10^5. On 2,500 miles of fiber, the round-trip delay is approximately 30 ms, giving a bandwidth $\times$ delay product of about 1.6×10^7 for an OC-12c circuit (for a 1.5 Mb/s satellite circuit, the bandwidth $\times$ delay product is about 7.7×10^5). Clearly, versions of TCP with high-speed modifications (extended windows, etc.) are needed to make effective use of high-speed ATM networks. Research is needed to determine whether current enhancements are adequate for the task.

Sequence number wrap: Under TCP, each octet of data is numbered. This *sequence number* is used by TCP's acknowledgement mechanism. On an OC-12c circuit transmitting at full speed, it would only take about 27 seconds for the 32-bit sequence number to wrap around and start reusing old values. A mechanism has been proposed to prevent this from causing problems [19].

Efficiency: Consider a null TCP data segment (e.g. an empty ACK). When this segment is transmitted over the ATM network, it consists of a (typically) 20-octet TCP header and a (typically) 20-octet IP header, so its length will be 40 octets. The AAL convergence sublayer occupies 8 octets (for either AAL 3/4 or AAL 5). If AAL 3/4 is used, the SAR takes 4 octets per cell, so even a null TCP segment doesn't fit in a single cell. If AAL 5 with no further encapsulation is used, a null TCP segment barely fits. If some further encapsulation (e.g. NLPID/SNAP) is used, no TCP segment fits in a single cell.

According to [8], about 40% of TCP wide-area traffic consists of empty ACKs; another 30% consists of TCP segments with 1–10 octets of data. These small segments end up being transmitted in two ATM cells, the second of which consists mainly of padding. This drives down the efficiency of the network. Various types of compression can improve efficiency; it remains to be seen whether this is really needed.

4.6 Performance

The speed that ATM networks are expected to attain makes performance a serious issue. In addition to the performance effects of TCP windows and transmission efficiency (discussed above), there is the issue of the load that the network can place on attached devices. At OC-12c speed (622.080 Mb/s, with 599.040 Mb/s available to the ATM layer), about 1.4 million cells per second are delivered to the UNI. Table 2 shows the *maximum transmission units* (MTUs) of various network media, along with the minimum and maximum IP packet sizes. For each packet size, the number of cells per packet and packets per second in the full OC-12c bandwidth are shown. Note that both AAL 3/4 and AAL 5 allow packets of the maximum size (65,535 octets).

Description	MTU	cells/pkt	pkt/s
Minimum	20	1	1,412,830
Generic	576	13	108,679
Ethernet	1,500	32	44,151
FDDI	4,352	91	15,526
HIPPI	65,280	1,361	1,038
Maximum	65,535	1,366	1,034
This table assumes AAL 5 encapsulation			

Table 2: Packet Sizes and Cell Rates

A common technique for improving communications performance is increasing the packet size. This reduces the number of interrupts that need to be handled and decreases the proportion of bandwidth used by overhead (packet headers, etc.). Using the maximum datagram size permitted by IP, a router only needs to handle a little over a thousand packets

per second. Router vendors can deliver products today that perform at this level, assuming that the ATM and AAL layers are implemented in hardware so there is not a per-cell demand on the router's CPU. It is, however, important to remember that traffic studies [7, 8] indicate that minimum-size packets make up a significant fraction of wide-area traffic.

4.7 Scalability

Today's experiments with ATM are being conducted on very small networks. As ATM is deployed throughout the network, these "islands" of ATM will grow and join, eventually evolving into a world-wide network. It is critical that the strategies, algorithms, and standards that are tested remain viable as they are scaled up to run over a global network.

5 Conclusion

ATM is an emerging technology that promises to figure heavily in both local- and wide-area networks. It is based on the transfer of fixed-length *cells* and allows for multiplexing through the use of *virtual circuits*. ATM is connection-oriented, but a proposed associated technology known as CBDS is connectionless. There are a number of unresolved issues associated with ATM, but quite a bit of work is being invested in it.

ATM networks need to interoperate, both with each other (e.g. LAN-WAN interoperation) and with networks based on other technologies (e.g. HIPPI, FDDI, Ethernet). While data-link-layer interoperation may be appropriate between networks based on ATM, network layer routing is more appropriate for interoperation between ATM and non-ATM networks. However, there are currently no commercially-available routers with adequate performance to handle routing between high-speed networks (e.g. a HIPPI LAN and a 622 Mb/s ATM WAN).

The introduction of a very-high-speed public wide-area network raises a number of issues for the IP family of protocols, including:

- internetworking strategies
- IP encapsulation
- optimization of routing
- IP address resolution
- connection management
- TCP considerations
  - cell loss
  - window management
  - sequence number wrap
  - efficiency
- performance
- scalability

Network research is needed to resolve these issues and to answer outstanding questions about running TCP/IP over ATM.

For the near term, these are our recommendations for building an IP internet which makes use of ATM:

Use AAL 5 as the ATM adaptation layer. It is designed to be efficient at transporting data, it provides excellent error detection, and it is the most likely AAL to become the eventual standard for data networking over ATM.

Pick one of the methods in [15] for encapsulation of IP datagrams. Any one of the mechanisms identified should work. One will eventually become the standard method, but it is impossible to say which one, so the network engineer should go ahead and pick one and make sure it is used throughout the network.

Use directed ARP for address resolution. If equipment cannot be modified to use directed ARP, the best alternative in the near term is probably hand-coded address resolution tables.

Use ATM permanent virtual circuits (PVCs). PVCs provide the most similar service to the fixed networks in use today, so using them should necessitate fewer changes than using switched virtual circuits (SVCs). In any case, most ATM equipment will not provide SVC capability in the short term.

Use a performance-enhanced TCP. It is imperative to implement the mechanisms in [18, 19, 20] to ensure good performance on high-bandwith networks.

A Acronyms

AAL	ATM adaptation layer
ACK	acknowledgement
ANSI	American National Standards Institute
ARP	address resolution protocol
ATM	asynchronous transfer mode
B-ISDN	broadband ISDN
BBN	Bolt Beranek and Newman, Inc.
BOM	beginning of message
CBDS	connectionless broadband data service
CCITT	International Telegraph and Telephone Consultative Committee
CL-PDU	connectionless protocol data unit
CLNAP	connectionless network access protocol
CLNAP-PDU	connectionless network access protocol protocol data unit
CLP	cell loss priority
CLSF	connectionless service function
COM	continuation of message
CPU	central processing unit
CRC	cyclic redundancy check
CS	convergence sublayer
CS-PDU	convergence sublayer protocol data unit
EOM	end of message
FDDI	Fiber Distributed Data Interface
FTP	File Transfer Protocol
Gb/s	gigabits per second
GFC	generic flow control
GOS	grade of service
HIPPI	High Performance Parallel Interface
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IP	Internet Protocol
ISDN	integrated services digital network
ISO	International Organization for Standardization
LAN	local area network
LLC	Logical link control
MB/s	megabytes per second
Mb/s	megabits per second
MID	message identifier
MTU	maximum transmission unit
NLPID	network layer protocol ID
ms	milliseconds
NNI	network node interface
OAM	operation and maintenance
OC-3c	Optical Carrier Level 3 Concatenated
OC-12c	Optical Carrier Level 12 Concatenated
OSI	Open Systems Interconnect
PBX	private branch exchange
PC	personal computer
PCI	protocol connection identifier
PDN	public data network
PDU	protocol data unit
PTI	payload type indicator
PVC	permanent virtual circuit
RFC	request for comments
SAR	segmentation and reassembly
SAR-PDU	segmentation and reassembly protocol data unit
SAR-SN	segmentation and reassembly sequence number
SDU	service data unit
SEAL	simple and efficient adaptation layer
SMDS	Switched Multi-Megabit Data Service
SMTP	Simple Mail Transfer Protocol
SNAP	subnetwork access protocol
SONET	Synchronous Optical Network
SSM	single-segment message
SVC	switched virtual circuit
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
UNI	user–network interface
VBR	variable bit rate
VC	virtual channel
VCI	virtual channel identifier
VP	virtual path
VPI	virtual path identifier
WAN	wide area network

B CCITT Recommendations for B-ISDN

I.113 Vocabulary of terms for broadband aspects of ISDN
I.121 Broadband aspects of ISDN
I.150 B-ISDN ATM functional characteristics
I.211 B-ISDN service aspects
I.311 B-ISDN general network aspects
I.321 B-ISDN Protocol Reference Model and its application
I.327 B-ISDN functional architecture
I.361 B-ISDN ATM layer specification
I.362 B-ISDN ATM Adaptation Layer (AAL) functional description
I.363 B-ISDN ATM Adaptation Layer (AAL) specification
I.364 Support of broadband connectionless data service on B-ISDN
I.413 B-ISDN user-network interface
I.432 B-ISDN user-network interface—Physical Layer specification
I.610 OAM principles of B-ISDN access

References

[1] ANSI T1S1.5/91-449. *AAL 5—A New High Speed Data Transfer AAL*. ANSI, November 4–8 1991.

[2] ANSI Working draft proposed American National Standard for Information Systems. *High Performance Parallel Interface—Memory Interface*. ANSI, March 21, 1991.

[3] ANSI X3T9.3/88-023. *High Performance Parallel Interface—Mechanical, Electrical, and Signalling Protocol Specification*, Revision 8.0. ANSI, 1988.

[4] ANSI X3T9.3/89-013. *High Performance Parallel Interface—Framing Protocol*, Revision 3.1. ANSI, 1989.

[5] ANSI X3T9.3/90-119. *High Performance Parallel Interface—Encapsulation of ISO 8802-2 (IEEE Std 802.2) Logical Link Control Protocol Data Units*, Revision 2.0. ANSI, 1990.

[6] ANSI X3T9.3/91-023. *High Performance Parallel Interface—Physical Switch Control*, Revision 1.7. ANSI, 1991.

[7] Cáceres, Ramón, Peter B. Danzig, Sugih Jamin, and Danny J. Mitzel. "Characteristics of Wide-Area TCP/IP Conversations," *Computer Communication Review*, Vol. 21, No. 4, September 1991, pp. 101–112.

[8] Cáceres, Ramón. "Efficiency of ATM Networks in Transporting Wide-Area Data Traffic," *Computer Networks and ISDN Systems*, in process.

[9] Cooper, Eric. *Transmission of IP Datagrams over Asynchronous Transfer Mode (ATM) Networks*. DRAFT, November 1991.

[10] Cypher, David and Shukri Wakid. "Standardization for SONET and ATM and Open Issues," *Proceedings of the 3rd Annual Workshop on Very High Speed Networks*, Greenbelt, MD, March 9–10, 1992

[11] Escobar, Julio, and Craig Partridge. "A Proposed Segmentation and Reassembly (SAR) Protocol for use with Asynchronous Transfer Mode (ATM)", *Proceedings of the 2nd IFIP WG6.1/WG6.4 International Workshop on Protocols for High Speed Networks*, Stanford, CA, November 1990.

[12] Garrett, John, John Hagan, and Jeff Wong. *Directed ARP*. Internet Draft, November 17, 1991.

[13] Goldstein, Fred R. *Compatibility of BLINKBLT with the ATM Adaptation Layer*. ANSI, T1S1.5/90-009. February 5, 1990.

[14] Händel, Rainer, and Manfred N. Huber. *Integrated Broadband Networks: An Introduction to ATM-based Networking*. Addison-Wesley. 1991.

[15] Heinanen, Juha. *Multiprotocol Interconnect over ATM Adaptation Layer 5*. Internet Draft, October 16, 1992.

[16] IEEE Std 802.2. *Information Processing Systems—Local Area Networks—Part 2: Logical Link Control*. IEEE, December 1989.

[17] ISO/IEC TR 9577. *Information Technology—Telecommunications and*

Information Exchange Between Systems—Protocol Identification in the Network Layer. ISO, October 1990.

[18] Jacobson, V. "Congestion Avoidance and Control". *Computer Communication Review*, Vol. 18, No. 4, August 1988, pp. 314–329.

[19] Jacobson, V., R. Braden, and D. Borman. *TCP Extensions for High Performance*. RFC 1323, May 1992.

[20] Jacobson, V., and R. Braden. *TCP Extensions for Long-Delay Paths*. RFC 1072, October 1988.

[21] Lyon, T., F. Liaw, and A. Romanow. *Network Layer Architecture for ATM Networks*. DRAFT, June 26, 1992.

[22] Piscitello, Dave, and Joseph Lawrence. *The Transmission of IP Datagrams over the SMDS Service*. RFC 1209, March 1991.

[23] Renwick, John K., and Andy Nicholson. *IP and ARP on HIPPI*. Internet Draft, January 1992.

[24] RFC 791. *Internet Protocol*. September 1981.

[25] SR-NWT-001763. *Preliminary Report on Broadband ISDN Transfer Protocols*, Issue 1. Bellcore, December 1990.

[26] Tsuchiya, P. *Shortcut Routing: Discovery and Routing over Large Public Data Networks*. Internet Draft, July 1992.