ATM: Paving the information superhighway

by R. A. Sultan
C. Basso

A new generation of networking requirements is fueling the growth of a cell-based communications technology known as the asynchronous transfer mode (ATM). ATM technology allows the integration of voice, video, traditional data, and other traffic types on a single network. ATM offers a unique opportunity to deploy the same standardized networking technology in both the wide-area and local-area environments. IBM has introduced a family of products that provides a complete ATM solution for customers. The products support ATM standards, allowing the products to interwork with devices from other vendors. This paper provides a tutorial on ATM technology and an overview of the IBM ATM product family. The IBM 8260 Intelligent Switching Hub is described as a representative ATM product.

The past few years have brought rapid growth to the Internet and a variety of private online computer services. Usenet and the World Wide Web have attracted a new community of casual network users. Increasing interest in personal videoconferencing, entertainment on demand, remote robotics, intelligent transportation systems, and other applications is placing new demands on communication networks. There has been a parallel growth in the popularity of multimedia computing, driven in part by the development of CD-ROM technology. Users increasingly expect the same high-quality multimedia service across a network as they receive locally. A proposed new communications infrastructure, known as the "information superhighway," is aimed at meeting these requirements. The concept of the superhighway is receiving wide attention in the press and by governmental organizations.

In addition to satisfying the requirements of emerging applications, there are strong economic incentives to aggregate traffic generated by existing applications. It is often the case today that distinct networks are constructed to provide specific types of service. Voice traffic is carried by switching equipment that provides relatively narrow bandwidth circuits having low delay properties. Data are often carried by packet routers that deliver large volumes of bursty traffic with less attention to delay properties. Substantial economies of scale can be realized by freely mixing these and other traffic types on the same switching equipment and communication links. This saving is most significant in wide area networks where the lease of trunk lines can represent a significant portion of the cost of network operation. Together, the requirements for emerging applications and for traffic integration suggest that networks of the future must provide traffic capacity and service quality tuned to the needs of individual applications.

Asynchronous transfer mode (ATM)[1] switching technology is widely viewed as the base for future global communications. It has been adopted as a

 0018-8670/95/$3.00 © 1995 IBM

standard for broadband communications by the International Telecommunications Union (ITU, formerly the CCITT) and by a communications industry consortium, the ATM Forum, which has over 700 member organizations worldwide. One key factor driving the growth of ATM technology is its integration of video, voice, and data traffic. Another is its support for guaranteed service quality and point-to-multipoint distribution capabilities needed by emerging network applications. A unique aspect of this technology is its applicability to both the wide-area and the local-area environments.

ATM technology relieves limitations on the geographic area that can be served by a local area network (LAN). The domain that can be served by an ATM LAN is often called a *campus*. The campus may be a collection of locally interconnected buildings, as in the case of a traditional university campus, or it may be a more widely dispersed collection of office buildings spanning a metropolitan area. In the campus environment, ATM offers significant advantages over conventional LAN technology. The current generation of LANs makes use of a shared transmission medium—a ring or bus. In contrast, ATM is a switching technology. A data packet transmitted through the switch need only contend for the availability of switch ports, rather than the availability of the entire medium. Data transfers can be performed in parallel across the switch, increasing the effective capacity of the medium. ATM networks can be increased in size by interconnecting additional switches, avoiding problems of scale encountered by conventional LANs. ATM technology also provides scalability with respect to speed, serving transport requirements in a range from a few megabits per second to gigabits per second. The convergence of wide-area and campus networking technology provides stronger guarantees of service quality when crossing network boundaries and simplifies network operation and management. ATM offers two dimensions of integration: aggregation of multiple traffic types in a single network and uniformity of network technology independent of geography.

It is anticipated that a wide variety of devices will be equipped for attachment to ATM networks, including video servers, public branch exchange (PBX) equipment, file servers, bridges, routers, and workstations. Application programs associated with these devices can be written to specify their service requirements to the network. It is likely that the sockets-based application program interfaces (API) in common use today will be extended for this purpose. The ATM-attached devices may contain the end users of the services, as in the case of workstations, or they may provide interconnection services for users in conventional networks, as in the case of ATM-equipped bridges and routers. Devices can also contain layers that shield existing programs from knowledge of the ATM network but allow them to benefit transparently from the network. An important example is LAN emulation. LAN services can be extended across an ATM network, offering higher throughput and lower latency than is offered by conventional LANs. The ATM Forum has specified a LAN emulation standard for this purpose.[2] The methodology allows existing LAN applications to use the ATM network without change and protects customers' investments in installed software. In addition to extending the geographic scope of LAN services transparently across the wide-area environment, LAN emulation allows the creation of distinct LANs based on work patterns or other administrative criteria, rather than physical connection to a medium. LANs specified in this way are called *virtual* LANs. The division of an ATM network into virtual LANs can be important in reducing the scope of broadcast traffic. Internet Protocol (IP) Address Resolution Protocol (ARP) requests, for example, are confined to a specific virtual LAN instead of being distributed across the entire ATM network. Virtual LANs can be interconnected by network routers, just as in the case of conventional LANs.

An ATM network can also be viewed as a subnetwork with respect to particular network layer protocols. Again, this methodology is transparent to application programs. The Internet Engineering Task Force (IETF) has, for example, specified a standard for the exchange of IP traffic over ATM.[3] IP-over-ATM is more efficient than IP-over-LAN emulation as it directs ARP requests to a server rather than broadcasting them. The scheme requires IP addresses and associated ATM addresses to be registered with the server.

The LAN emulation and IP-over-ATM schemes allow existing applications to make use of the ATM network, but they do not make good use of the traffic and quality guarantees offered by ATM. There is currently an effort by the IETF to define a future version of IP, known as *IP Next Generation* (IPNG),[4] which allows specification of quality-of-service (QOS) parameters associated with a traffic flow. ATM-equipped IPNG routers and hosts can use the

ATM network to satisfy the QOS requirements of a flow. The variety of alternatives for the use of ATM networks, ranging from transparent access by existing application programs to the use of an ATM API by new application programs, provides a useful staging for the introduction of ATM networks.

The growth of ATM has benefited from a strong standardization effort. A specification for the attachment of devices to ATM networks has been completed.[5] There is an interim standard for the interconnection of switches within an ATM network.[6] Products conforming to these standards are now available. A further specification for the interconnection of ATM switches, and the hierarchical organization of switches within an ATM network, is now being completed.[7] Standardization allows customers to construct ATM networks using equipment from a variety of vendors. Standardization provides vendors with a clear set of specifications on which to base products. It can have the effect of speeding new products to the marketplace. Although ATM interfaces are standardized, device internal technology is not. Thus, ATM products can be differentiated on the basis of price, performance, capacity, usability, serviceability, reliability, and other factors. Standard ATM interfaces can be supported at the boundaries of a network whose individual switches are interconnected via proprietary interfaces. Such a network can provide added levels of performance and function while still providing ATM standard interfaces to users.

IBM has announced a comprehensive high-speed networking strategy based on ATM technology. New products include ATM switches for the campus and the wide area network, adapters and software for attachment to the network, and equipment for interconnection between conventional and ATM networks. The product family is depicted in Figure 1. Attachment to the network is provided by IBM Turboways* ATM adapters. A 100-megabits-per-second (Mbps) version provides high-speed fiber optic attachment of devices directly to the campus ATM switch. A 25-Mbps version allows devices to be attached to a workgroup concentrator via twisted-pair wiring. The use of twisted-pair wiring permits customers to maintain their existing wiring infrastructure while migrating to ATM technology. The IBM 8282 Workgroup Concentrator is, in turn, attached to the campus switch via 100-Mbps fiber optic cable. The workgroup concentrator is not an ATM switch. It cannot route ATM traffic from one workstation to another, but instead, aggregates relatively low-speed ATM traffic before sending it to the campus switch. Aggregation of traffic by the workgroup concentrator makes it possible to provide low-speed ports, suitable for entry-level ATM users, at a low cost. LAN emulation software is available for use with the Turboways adapters, allowing users to view the ATM network as a virtual LAN.

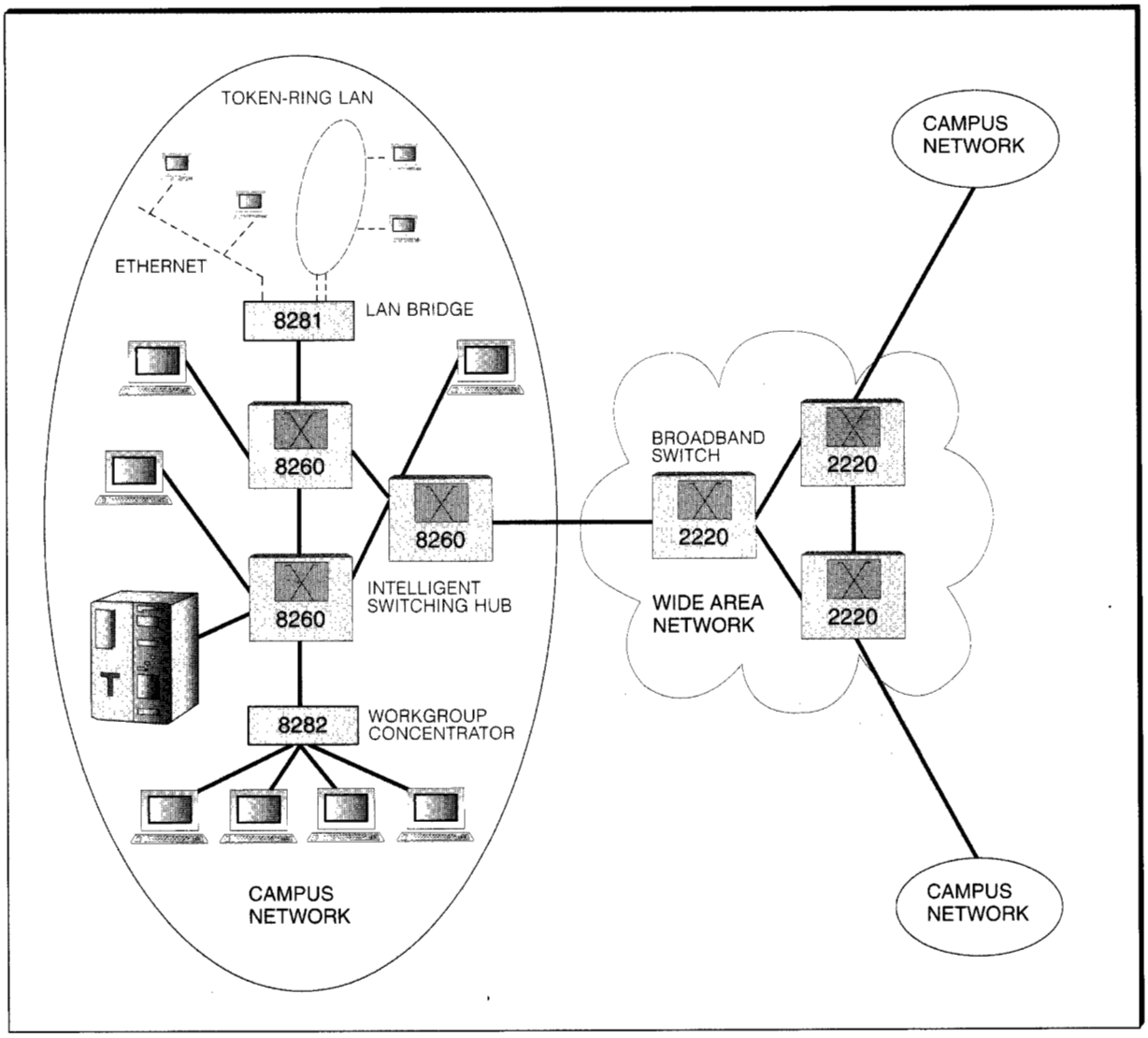
ATM growth has benefited from a strong standardization effort.

The IBM 8281 ATM Bridge provides interconnectivity between conventional LANs and ATM virtual LANs. The bridge uses LAN emulation protocols to communicate across the ATM network but, in other respects, behaves as a conventional LAN bridge, participating, for example, in bridge protocols. The IBM 8260 Intelligent Switching Hub, or simply, the Hub (developed and marketed jointly by IBM and the Chipcom Corporation), provides both the functions of a campus ATM switch and a conventional LAN wiring concentrator. The packaging allows users to migrate gradually to ATM function by replacing LAN ports with ATM ports. The customer's investment in the base Hub equipment is preserved during this evolution.

The initial ATM product for the wide area network is called the IBM 2220 Broadband Switch. The broadband switch has been designed to support large-capacity trunk lines at high levels of utilization. The switch supports IBM's Networking BroadBand Services (NBBS) that offer a variety of functions not found in standard ATM networks. The broadband network supports ATM interfaces at its boundaries, allowing interconnection of campus networks across the wide area network.

In this section we have outlined the strategic importance of ATM technology and described IBM's emerging ATM products for the campus and the wide area network environments. In the next sec-

Figure 1 ATM networking products



tion we examine ATM technology and standards in more detail, and in a following section, we see how this technology has been applied to the campus ATM switch.

ATM cell-based transport

ATM is a connection-oriented switching technology in which traffic is carried in fixed-size packets called *cells*. Connections are either point-to-point, with bidirectional traffic flow between endpoints, or point-to-multipoint, with unidirectional flow from a single root to a set of leaves. Connections are created with the properties needed by specific traffic streams. These properties are expressed by traffic and QOS parameters. Traffic parameters describe, among other things, the capacity, or cell rate, of the connection. Traffic parameters can be specified separately for each direction of data flow in the case of point-to-point connections. Quality of service is described by a set of parameters related to the speed, accuracy, and availability of the

connection. One such parameter is a bound on the delay experienced by traffic. Bounded delay is important in time-critical applications such as robotics and in applications such as voice conversation where excessive latency can inhibit usability. Another parameter, the variation in delay, known as *jitter*, is important for real-time audio and video distribution applications. Excessive jitter can distort the playback of audio and video streams. In some cases the traffic and QOS requirements of streams may not be known or may vary widely over time. It is possible to specify traffic and QOS parameters on the basis of an estimate of peak requirements. Such a policy, however, may be wasteful of network resources required to provide the guaranteed service levels. An alternative is to specify connections as providing available bit rate (ABR) service, in which case traffic capacity and QOS are not guaranteed.

An ATM network is made up of the trunk lines that carry traffic and the switches that move traffic from one trunk line to another. The switches are responsible for guaranteeing the connection properties requested by users. The cells that carry ATM traffic are uniform in size and are relatively small (53 bytes, including a five-byte header). Cells associated with various connections are interleaved on the trunk lines and switches of the network. Cells with specific traffic and QOS requirements can receive appropriate priority as they travel through the network. Such prioritization schemes are beyond the scope of the ATM standards and are a proprietary feature of vendor switches. It is characteristic of ATM, as an *asynchronous* transfer mode, that cells, after prioritization, can be forwarded as soon as capacity is available for their transmission. Cells are not required to wait for a particular slot, or time allocation, as is generally the case with *synchronous* transfer modes. The fixed-cell size allows a bound to be placed on the maximum delay that a high-priority cell encounters while waiting for a cell transmission in progress to complete (without preemption of a cell transmission in progress). The small size of cells ensures that the delay bound is low. It is also easier and less costly to design switching hardware for cells of a uniform size. The ability of ATM technology to effectively integrate a variety of traffic types in a single network rests largely on its use of small cells that can be interleaved and prioritized.

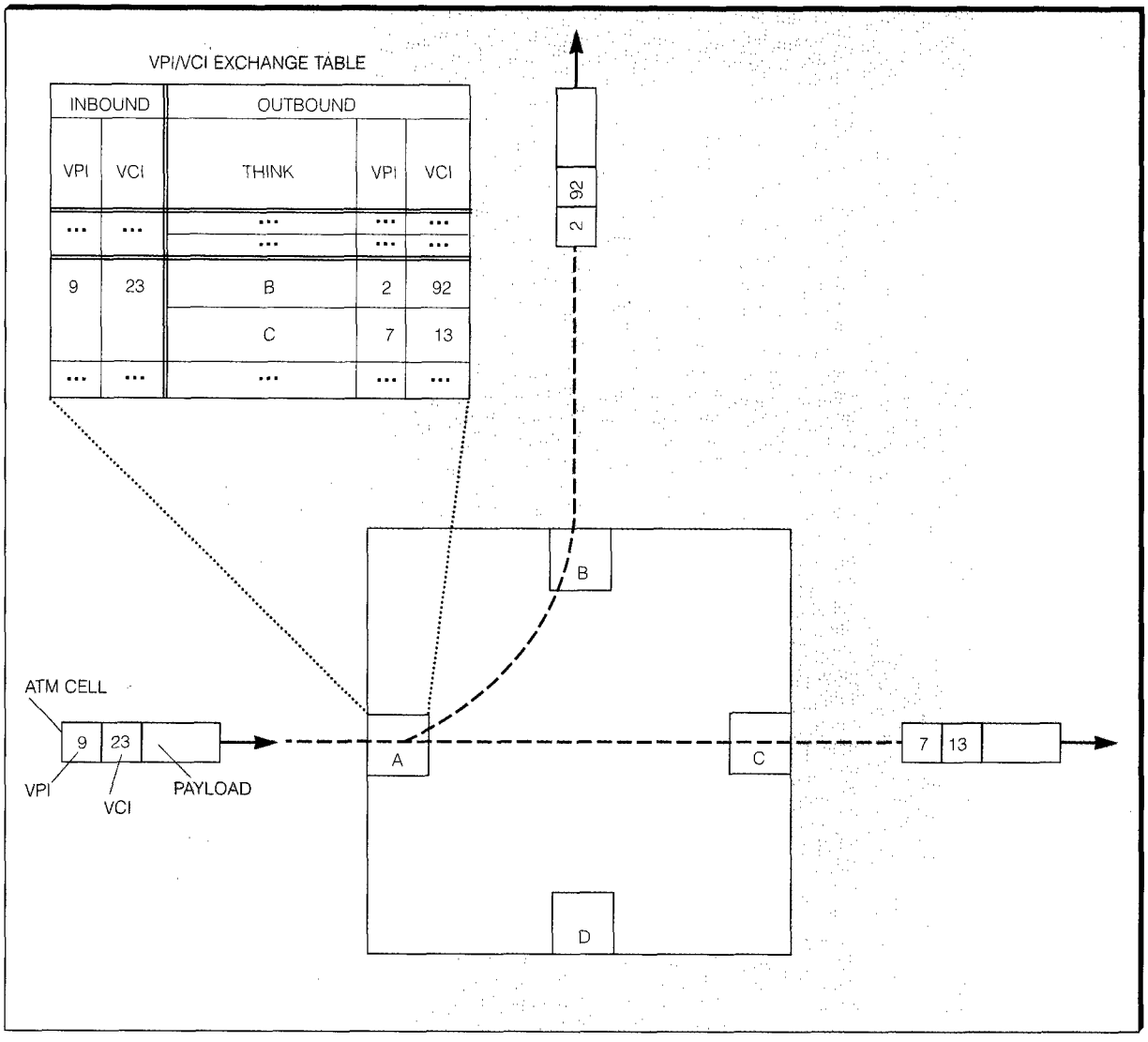
Users of ATM connections are located in end systems attached to the periphery of the ATM network by an access trunk line. The interface between the end system and the switch to which it is attached is called the *User-to-Network Interface* (UNI). The interface used for the interconnection of switches within the ATM network is called the *Network Node Interface* (NNI). The UNI and NNI specifications form the core of ATM standardization. An ATM connection is composed of a sequence of concatenated segments. A segment is that part of a connection associated with one particular trunk line. Each segment supports the QOS and traffic parameters required end-to-end by the connection. A segment associated with a UNI is identified by a three-byte *VPI/VCI* value. The VPI/VCI value is drawn from an address space associated with the trunk line. The VPI/VCI value is composed of a one-byte virtual path identifier (VPI) and a two-byte virtual channel identifier (VCI). The NNI allows a 12-bit VPI, with a correspondingly larger VPI/VCI value.

Figure 2 shows an ATM cell, inbound to the switch, carries the VPI/VCI value of the segment on which it arrives. This inbound VPI/VCI value is used as the key in a search of a VPI/VCI table associated with the inbound trunk line. In the case of a point-to-point connection, or a point-to-multipoint connection that does not branch at the particular switch, the target entry contains the identity of the outbound segment. The switch forwards the cell on the outbound segment, replacing the inbound VPI/VCI value with the outbound value found in the table. In the case of a point-to-multipoint connection that branches at the particular switch, the target table entry identifies a list of outbound segments and associated VPI/VCI values. The switch forwards the cell on each of the outbound segments, replacing the inbound VPI/VCI value with the outbound VPI/VCI value associated with the segment. This process of VPI/VCI exchange and cell forwarding is the fundamental method of transport in ATM networks.

The ATM connections discussed thus far are known as virtual channel connections. Multiple virtual channel connections can be bundled on a single virtual path connection. Cells transported on a virtual path connection are switched using only the VPI portion of the VPI/VCI value.

ATM connections provide a cell transport service. ATM users, however, require services more specific to their needs. An ATM adaptation layer (AAL), provided at the endpoints of ATM connections, supplies such application-specific services. A variety

Figure 2 VPI/VCI exchange and ATM cell forwarding



of AAL types have been defined. AAL1, for example, supports constant bit rate (CBR) transmission as might be required by a voice conferencing application. AAL5 supports the transmission of variable length frames used in conventional LAN environments. End users obtain their services from an AAL, and the AAL, in turn, uses the cell transport service provided by ATM connections. The communication among AAL users is known as a *call*, and the users are said to be *parties* to the call. Current standards specify that a call is mapped one-to-one to its underlying ATM connection, and for many purposes it is not necessary to distinguish the two. Future specifications may allow the multiplexing of calls on a connection.

In this section we have described the fundamentals of ATM cell transport and the method by which the AAL uses cell transport in order to provide higher-level services to users. These activities occur in steady state during the operation of an ATM network and are said to belong to the ATM user plane.

Equally important are the activities required to establish the steady-state environment. Such functions are performed by the control plane. A third set of functions, performed by the management plane, allows access to information associated with the ATM network and end systems. In the next section, we will examine the role of the control and management planes in the establishment of ATM connections. This role has been a key focus of standardization efforts.

Establishing ATM connections

ATM connections are said to be either permanent or switched. Permanent connections are initiated via an operator or management interface by specifying the endpoints and characteristics of the connection. Permanent connections are generally used when end systems do not have signaling capabilities or when the network operator wishes to have a high degree of control over the establishment of connections. A special case of permanent connections is that of connections whose VPI/VCI values are reserved for use by ATM control and management functions. The establishment of such reserved connections occurs implicitly, for example, when trunk lines become active. There is currently no standard methodology for the establishment of permanent connections. Switched connections are initiated by the user applications themselves via signaling protocols between the end system and the network. Switched connections generally make better use of network resources than permanent connections since they are established when needed. They are more likely to reflect the specific requirements of user applications since they are initiated by the applications. For these reasons, the focus of standardization efforts has generally been on switched connections.

The establishment of a switched connection requires that the endpoints of the connection, the end systems, have an address that is known to both end system and network. ATM end systems are identified by the first 19 bytes of a 20-byte ATM address. The first 13 bytes are called the *prefix* and identify the switch to which the end system is attached. The next six bytes, called the *end system identifier* (ESI), uniquely identify a particular end system among end systems associated with the switch. It is necessary for both the end system and the network to know the full 19-byte value that uniquely identifies the end system. The ATM Forum defines an interim local management interface (ILMI) to allow the end system and access switch to exchange ESI and prefix values. The exchange occurs when both end system and switch have become active. The ILMI uses the Simple Network Management Protocol (SNMP) over a reserved ATM connection in order to perform the exchange. After the ILMI exchange has been performed, the end system can signal the network in order to establish connections.

The UNI signaling protocol[5] is based on an ITU standard known as Q.2931. The exchange of signaling messages drives transitions in finite state machines associated with the user and network sides of the UNI. Transitions are accompanied by actions associated with the setup of the connection. These actions include the establishment of appropriate VPI/VCI table entries as well as traffic and QOS parameters that characterize the connection. State transitions also drive the forwarding of signaling messages to other switches or end systems and the reply to signaling messages. The signaling protocol is performed using a connection, called the *signaling channel*, having the reserved value VPI = 0 and VCI = 5. Signaling messages are delivered reliably across the signaling channel using an extension of AAL5, called the *signaling AAL* (SAAL). AAL5 itself provides no guarantee of message delivery. The SAAL performs functions such as sequence number checking and retransmission.

A signaling message is identified by a message type and contains a set of information elements specific to the message type. Examples of signaling messages are the SETUP, which initiates establishment of the connection; CONNECT, which indicates to the sender of the SETUP that the connection has been established; and ADD PARTY, which requests that a party be added to an existing connection. UNI signaling provides the method by which users request the establishment of connections. It is also the protocol for the establishment of connection segments associated with the UNI. UNI signaling does not, however, specify the means by which connection segments are established within the ATM network. This function is performed by NNI signaling. The ATM Forum specifies a standard for NNI signaling known as the Private Network Node Interface, or P-NNI. The P-NNI provides protocols for the establishment of connection segments, similar to the UNI, but also provides a methodology for specifying the path of the connection. The P-NNI has been developed in two stages, known as phase 0 and phase 1. Phase 0 is specified by an Interim

Inter-Switch Signaling Protocol (IISP).[6] The IISP uses a slightly modified version of the UNI signaling protocol to establish calls across the NNI. The IISP provides a simple scheme to specify the path of a call. Each switch contains a table listing ATM address prefix values. When processing an arriving SETUP message, the table is searched for a prefix matching that of the destination ATM address. The matching entry specifies the interface on which the SETUP is forwarded.

The Hub is a campus ATM switch that implements the network side of the UNI signaling protocols.

The P-NNI phase 1[7] uses a signaling scheme that differs somewhat from the IISP. Both of the interconnected switches use the network-side protocols so that P-NNI phase 1 signaling is symmetric. This use makes it unnecessary to arbitrarily assign roles to the switches. A more significant difference between phase 0 and phase 1 is that phase 1 provides a multilevel hierarchical routing model, replacing the simple forwarding table described by the IISP. At the lowest level of the hierarchy, switches are grouped into clusters called *peer groups*. Switches exchange information in order to construct a database, replicated at each switch, describing the topology and state of the switches and trunk lines within the peer group. This methodology is well known and is commonly called the *link state advertisement*. The database is used to compute paths with, for example, a Dijkstra shortest path algorithm.[8] The path takes the form of a designated transit list (DTL) that is carried as an information element in the SETUP message. At each switch, the DTL is parsed in order to determine the switch to which the message should be forwarded. At the next level of the hierarchy, the lower-level peer group is viewed as a logical node within the higher-level peer group. State information exchanged by second-level logical nodes represents a summary or abstraction of first-level information. The scheme is repeated to provide additional levels of routing hierarchy.

In this section, we have described key standards associated with the establishment of ATM connections. These standards include the ILMI exchange required to establish the ATM addresses, the Q.2931 and SAAL protocols associated with UNI signaling, and the two phases of P-NNI signaling. In the next section, we will take a detailed look at IBM's ATM campus switch, the 8260 Intelligent Switching Hub. We will describe both its support for ATM standards and some important functions unique to the product.

The 8260 ATM Hub

The Hub is a campus ATM switch that implements the network side of the UNI signaling protocols. Conformance to the UNI standard ensures that an end system of any manufacture, implementing the corresponding user-side standards, can use the services of the campus ATM network. The Hub implements a two-level NNI routing hierarchy. The first-level NNI, called the switch-to-switch interface (SSI), uses a scheme similar to the P-NNI phase 1. The P-NNI phase 1 specification was not available at the time of development, so the SSI is an approximation to the standard. Later versions of the Hub will provide full compliance to the P-NNI phase 1. The second level of the NNI hierarchy conforms to the ATM Forum P-NNI phase 0, allowing the Hub to interconnect with hubs from other venders that conform to the P-NNI phase 0 standard. In the future, all levels of the NNI hierarchy will conform to P-NNI phase 1, but it is likely that support for P-NNI phase 0 will be maintained for compatibility with products not supporting phase 1.

In addition to local interconnection via fiber optic cable, Hubs can be interconnected using the services of a wide area network, as might be provided by a collection of IBM 2220 Broadband Switches. In the case of wide area network (WAN) interconnection, a single Hub port provides interconnection with multiple remote campus networks. The channel interconnecting a pair of Hubs across the WAN is an ATM permanent virtual path connection, supplied by the WAN (the only type of interconnection generally offered by ATM bearer services today). ATM connections that are bound from one campus to another, across the WAN, are bundled on the appropriate virtual path. The virtual paths

are created using traffic parameters consistent with expected intercampus traffic. Establishing ATM connections across a virtual path connection, rather than a trunk line, requires some change to the NNI signaling protocols. In particular, signaling protocols usually performed over a reserved channel associated with VPI = 0 use instead a channel associated with the interconnecting virtual path.

The Hub supports both reserved-bandwidth and available bit rate connections. Reserved-bandwidth connections guarantee peak traffic requirements. Reserved-bandwidth traffic entering the Hub is policed to verify that it does not exceed the established traffic parameters for the connection. Traffic is shaped by respacing cells as they pass through the Hub. Shaping reduces the effect of variations in switching delay and prevents excessive jitter from accumulating along the path of the connection. Sufficient storage is allocated for each such connection so that cells associated with policed and shaped traffic are not lost or dropped while awaiting transmission through the switch. ABR connections do not explicitly reserve bandwidth but make use of available capacity. A basic level of ABR service is ensured by statically reserving approximately 25 percent of trunk bandwidth for this purpose. Support for ABR traffic is particularly important during the early stages of ATM deployment. Schemes such as LAN emulation and IP-over-ATM use ABR service because of the relatively unpredictable nature of the traffic they carry. The ATM Forum UNI standard specifies a rate-based end-to-end flow control scheme that allows the receiver to specify the rate at which the sender can transmit. The rate-based scheme differs from traditional window flow control schemes in that it specifies the rate at which cells can be received rather than the number of cells that can be received. Congestion *within* the ATM network can still, however, cause loss of cells, since there is no reservation of resources associated with the ABR traffic. It is assumed that users of ABR service recognize the possibility of delivery failure and retransmit when loss occurs. Retransmission can, however, result in spiraling levels of traffic. Loss of a single cell might, for example, result in the retransmission of a frame containing 20 cells.

The Hub provides an ATM management plane that supports a number of standard management information bases.

While maintaining the appearance of the ATM Forum standard scheme at the UNI, a network of interconnected Hubs implements a proprietary hop-by-hop flow control scheme that reduces the likelihood of ABR cell discard. A Hub determines, based on a queue occupancy threshold, when it has become congested. It notifies the upstream Hub that is the source of the traffic to halt its flow. If the source continues to send cells, the upstream Hub will also become congested and will notify its predecessor to halt traffic. Eventually, notification propagates to the traffic source. This method of hop-by-hop flow restriction is known as *backpressure*. An important feature of the backpressure scheme of the Hub is that it operates on individual connections. Traffic on well-behaved connections continues to flow while traffic is halted on connections causing the congestion. Only when the aggregate queue occupancy threshold is crossed is traffic halted on a trunk. The efficient support of ABR traffic is an important feature of the Hub.

The Hub provides an ATM management plane that supports a number of standard management information bases (MIBs), which are accessed using SNMP. The supported MIBs include MIB II,[9] which provides a set of base management objects found in many networking environments; the IETF AToMMIB,[10] which provides objects specific to the ATM environment; and the OSPF (Open Shortest Path First) MIB,[11] which provides objects associated with network topology and route computation. Some Hub-specific MIB extensions are also supported. The Hub contains a network management agent that communicates with an external network management application, such as Netview/6000*. In order to support such communication, the Hub implements IP-over-ATM protocols.

In addition to its standard UNI signaling support for establishing ATM connections on demand, the Hub also allows the specification of permanent connections via the network management applica-

tion interface. Users need not specify the exact path of a permanent connection. The Hub computes the path associated with a permanent connection and reuses much of the switched connection signaling support in establishing permanent connections. Configuration of the Hub is performed via a terminal dialog interface that can be accessed locally using an attached ASCII display or remotely via a TELNET session. Configuration information is retained in the flash memory of the Hub so that subsequent startup does not require operator intervention.

In this section, we have seen how the IBM 8260 Hub provides both conformance to standards *and* features that distinguish it in the marketplace. Examples of the former are its adherence to the UNI and P-NNI phase 0 protocols and use of IP-over-ATM and SNMP standards. Examples of the latter are the use of selective backpressure for improved ABR performance, WAN connectivity utilizing virtual path connections, an efficient methodology for the establishment of permanent connections via a network management interface, and the combination of conventional LAN hub functions in the same packaging as the ATM switch. Having established a functional view of the Hub, we now take a brief look at how these functions are realized in the physical structure of the product.

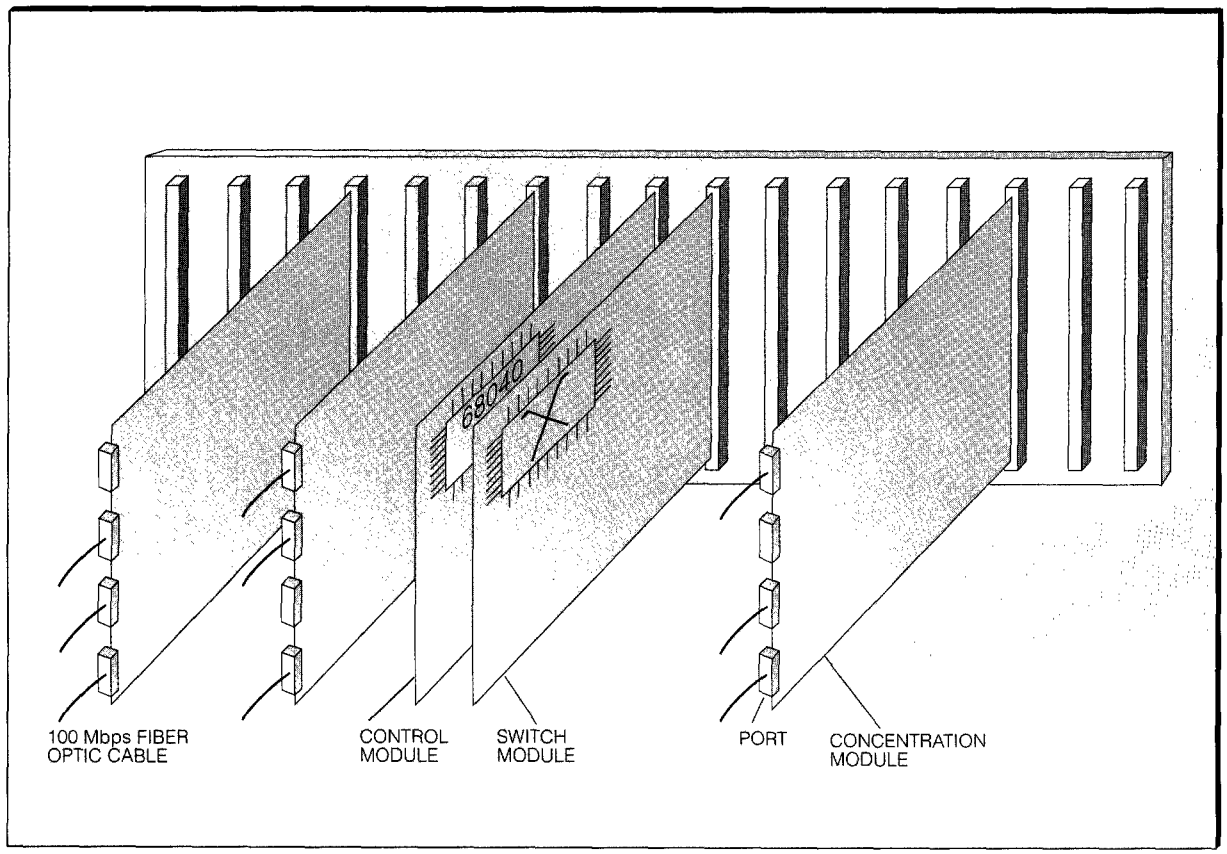
Inside the Hub

At the core of the Hub is a 16-port switch fabric on a CMOS VLSI chip.[12] A port of the switch is called a *leg* to distinguish it from the external ports of the Hub. This "switch-on-a-chip" technology is the same as that used in the IBM 2220 Broadband Switch. Data packets introduced at an input leg are delivered to one or more output legs. The switch provides bandwidth of approximately 270 Mbps from leg to leg in each direction of data flow, yielding an aggregate duplex capacity of approximately 4.3 Gbps. As can be seen in Figure 3, the Hub has a backplane with 17 vertical slots for the insertion of modules of various types. One slot is reserved for a switch module housing the switch chip and a switch control processor. A second slot is reserved for a control module containing a Motorola 68040 processor that executes the ATM control and management functions. The switch and control modules are tightly coupled and are plugged into adjacent slots in the backplane. Other slots are populated by concentration modules. Each concentration module provides four ports supporting external line attachment via 100-Mbps fiber optic cable. The port implements a standard called the *Transparent Asynchronous Transmitter-Receiver Interface* (TAXI), which is one of a number of ATM Forum standard physical interface specifications. Future releases of the Hub will support the Synchronous Optical Network (SONET) interface specifications to provide higher-capacity trunk line interconnection. A port can be configured to support a UNI, SSI, or P-NNI phase 0. The port can also be associated with multiple virtual path connections in the case where it is attached to an ATM WAN.

The control module and concentration modules are wired directly to legs of the switch and communicate with one another through the switch. ATM cells traveling on established ATM connections through the Hub are processed entirely by the hardware of the concentration and switch modules. The switch module transports ATM cells from an inbound concentration module to one or more outbound concentration modules. The concentration modules perform ATM user plane functions such as VPI/VCI swapping. When an ATM cell arrives at an inbound concentration module via an external port, the VPI/VCI value is extracted, and the VPI/VCI table is searched to find an entry representing the connection. If the cell is to be switched to a single outbound concentration module, the table entry supplies a target module number. If the cell is to be sent to multiple modules, the entry contains a seven-bit multicast identifier. The module number or multicast identifier is placed in a header that is prepended to the cell as it is sent to the switch. When the cell arrives at the outbound concentration module, the VPI/VCI table is again searched to find an entry representing the connection. The entry indicates the outbound port or ports of the concentration module to which the cell is delivered.

Most ATM control functions are performed in the control module. The VPI/VCI tables are, however, located in the concentration modules. Information needed to maintain the swap tables is sent from the control module to the concentration modules in control cells transported via the switch. Control and management information is, in general, transported from module to module via such control cells. The scheme makes it unnecessary to implement distinct wiring for control functions and simplifies the design of the hardware.

Figure 3 Hub ATM hardware structure



We have described the basic organization of the Hub. The next section examines how the Hub has introduced some of the important features of the P-NNI phase 1 before the availability of the standard. These features include the computation of connection paths, the distribution of network topology information, and the hierarchical organization of the ATM network. Although standards-compliant versions of these functions will eventually be available, the Hub provides an approximation to these functions in its initial release.

A hierarchical network organization

Hubs are organized into a two-level hierarchy. At the lower level of the hierarchy, individual Hubs are grouped into a cluster. Hubs within a cluster are interconnected by the SSI, which approximates the first level of the P-NNI phase 1 specification (an interim solution pending implementation of P-NNI phase 1). The SSI includes switch-to-switch signaling, a specification of the path of a connection, and a link state protocol used in the maintenance of intracluster topology and state information. Up to 256 clusters can be interconnected to form an ATM network. Hubs at the boundary between clusters implement the P-NNI phase 0 protocols on the intercluster interface. Collections of non-IBM hubs can be viewed as clusters, provided they implement the P-NNI phase 0 protocols. In addition to providing a means of interconnecting equipment of different manufacture, clustering is also useful in reducing the scope of network control traffic. Lowest-level topology information, for example, is not propagated from cluster to cluster. It is particularly important to avoid such traffic across the wide area network, where bandwidth is a costly resource. Clustering provides a means to isolate

control traffic to the local campus when interconnection is provided across a WAN.

In order to determine the path to a given ATM end system, it is first necessary to identify the Hub serving that end system. Recall that ATM addresses are constructed via an ILMI exchange in which the Hub supplies the prefix portion of the address. This scheme allows the Hub to encode information about itself in the portion of the ATM address that it supplies. The first 11 bytes of the 13-byte prefix are identical for all Hubs within the network. The twelfth byte contains a number uniquely identifying the cluster to which the Hub belongs. The thirteenth byte contains a number uniquely identifying the hub within the specified cluster. Although this particular partitioning of the ATM address is specific to IBM Hub networks, the scheme is consistent with the use of ATM addresses described in the P-NNI phase 1 standard. In addition to the prefix value, the operator also specifies a single ESI that identifies a virtual end system associated with Hub internal users. The virtual end system can be visualized as containing, for example, the network management agent that communicates via ATM connections with entities external to the Hub. The scheme of imbedding hub and cluster numbers in the ATM address allows a trivial mapping from a given ATM address to the identity of the Hub associated with that address. It was noted earlier that ATM switches of other manufacture can be represented by a cluster. These switches, in general, do not implement the Hub's scheme of imbedding hub and cluster numbers in the ATM address. Such addresses can be recognized because the first 11 bytes do not match the single value used throughout the Hub network. These "foreign" addresses are resolved by reference to a table associating ATM addresses with hub and cluster numbers.

Hub intracluster routing, like P-NNI phase 1 routing, assumes path computation based on a view of intracluster topology that is replicated in each Hub belonging to the cluster. In order to maintain such replicated topology information, the Hub has reused, with some modification, a link state algorithm frequently used in router environments called *Open Shortest Path First* (OSPF).[13] The message used by OSPF to carry information about the state of a link, is called a *link state advertisement* (LSA). The OSPF algorithm has been enhanced so that LSAs reflect the bandwidth currently available on intracluster trunk lines. In addition to describing intracluster trunk lines, the topology database contains the intercluster trunk lines that can be used to exit the cluster.

OSPF, having been developed for the router environment, is based on the use of IP addressing. The assignment of IP addresses to trunk lines is performed dynamically and is hidden from users of the Hub. IP address assignment does, however, require that information, such as hub numbers, be exchanged between Hubs at the time of trunk-line activation. This exchange of information on the SSI is known as the node identifier exchange (NIX).

In the router environments where OSPF is generally used, LSAs are generally issued when there is a change in network connectivity, a relatively infrequent event. The Hub, however, uses LSAs to distribute changes in available bandwidth. Without some form of damping, an LSA could be generated each time an ATM connection segment is established. The processing and bandwidth requirements of such frequent distribution could be excessive. For this reason, an LSA is issued only when available bandwidth has changed by more than a threshold value, currently set to 5 percent of the trunk bandwidth. OSPF itself provides additional damping by limiting the frequency of LSA generation to approximately once-per-second per trunk line.

In this section, we have described how OSPF has been reused to provide a method of topology and state distribution for ATM networks. Now we examine how OSPF routing has been similarly adapted.

OSPF-based path computation

OSPF is used by the Hub to compute paths within a cluster. If the destination lies outside the cluster, the path is computed to the trunk line by which the connection exits the cluster. In this case, independent path computations are performed in each subsequent cluster. The topology database, constructed via the distribution of LSAs, can be viewed as a graph of the network. In the case of the Hub, edges represent intracluster trunk lines and intercluster links exiting the local cluster. In router environments, OSPF is generally used to establish next-hop tables. A next-hop table entry indicates the local IP interface instance on which an arriving packet should be forwarded in order to reach a specified destination. The Hub intracluster routing methodology, like the P-NNI phase 1

methodology, computes a path at the connection originating Hub. This method is called *source routing*. The path is described by a list of path segments, called a *route vector*. The route vector is similar to the designated transit list of the P-NNI phase 1 specification. OSPF has been extended to provide the computation of route vectors.

OSPF uses a Dijkstra shortest path algorithm to compute routes. The route to a destination is the one with the lowest sum of edge weights. In an ATM network, it is a requirement to provide a route that supplies a requested cell rate. For this reason, it is necessary to exclude edges that do not provide sufficient bandwidth. When adding parties to an existing point-to-multipoint connection, any edge already included in the connection can be marked as having an infinitesimal cost. (The algorithm does not work properly if each added edge does not increase the weight of the path by some amount.) This modification of the shortest path algorithm is used by the Hub for the computation of routes associated with point-to-multipoint connections. The same method could be used to compute paths for point-to-point connections, but a method called *widest-path computation* is used instead in order to reduce the time required to establish connections.

For point-to-point connections, it is possible to precompute a tree of widest paths to destinations.[14] This method facilitates rapid establishment of point-to-point connections since no computation is required at the time connection is established. Consider all possible paths from the connection origin to destination. Each path has one edge (or several edges of equal bandwidth) that is the narrowest edge in the path with respect to available bandwidth. The bandwidth provided by the path can be no larger than that provided by this smallest, or bottleneck, edge. The widest path algorithm chooses the path having the bottleneck edge with the largest available bandwidth. This choice provides a path having the largest available end-to-end bandwidth to the destination. The computation is performed each time there is a significant change in available bandwidth associated with an edge. Such computations can be performed in the background and, in general, do not add to the latency of establishing a connection. The method guarantees a precomputed path of sufficient bandwidth from origin to destination, provided that such a path exists. When a suitable path does not exist, the connection request is denied. The penalty for precomputation is that the path may not be optimal with respect to the global allocation of bandwidth.

An example of widest path routing is shown in Figure 4. The widest path from A to D is A-B-C-D. The smallest, or bottleneck, edge in this path is A-B. This bottleneck is larger than the bottleneck of any alternative paths from A to D (circled numbers indicate the width, or bandwidth, associated with the edge).

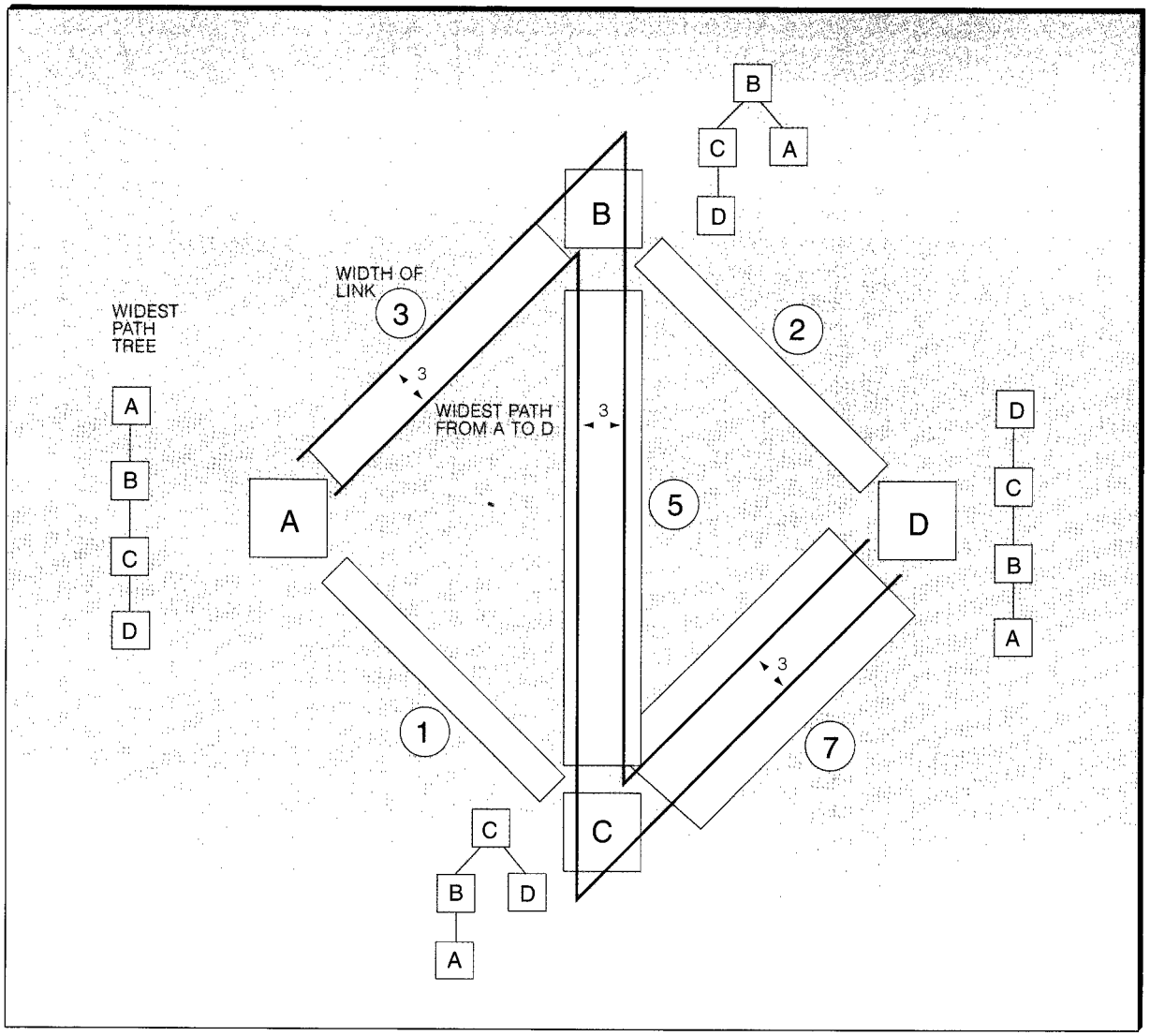
Summary

A family of emerging ATM products provides customers with a uniform broadband networking solution across both the campus and wide-area environments. The products offer standard interfaces that allow interworking with products of other vendors. They also offer unique features, such as the high level of trunk-line utilization provided by the broadband switches. The 8260 Intelligent Switching Hub is IBM's initial entry in the emerging market of ATM campus switches. The Hub switching technology supports an aggregate throughput in excess of 4 Gbps. By combining the functions of a conventional LAN hub with the functions of an ATM switch, customers can migrate gradually to the ATM environment. The Hub supports ATM Forum UNI and P-NNI phase 0 standards. Functions associated with P-NNI phase 1 have been provided prior to the availability of the completed standard. It has been accomplished by the adaptation of OSPF routing to meet ATM requirements.

Acknowledgments

The authors would like to acknowledge the work of their colleagues who participated in the development of the IBM 8260 ATM Hub. Contributors from the IBM Networking Hardware Division, La Gaude, France, are: Marianne Aubry, Annette Beaulieu, Jean Calvignac, Paul Chemla, Philippe Damon, Jean-Yves Demachy, Tri Dokhac, Eric Forestier, Jean-Francois Gilbert, Laurent Gimet, Mathieu Girard, Dominique Godard, Claire Lamy, Guy Menanteau, Elliott Norsa, Daniel Orsatti, Stephan Pacchiano, Jean-Bernard Schmitt, Paulo Scotton, Michel Susini, Fred Truco, Sylvie Ubaldi, and Fabrice Verplanken. Contributors from the IBM Research Division, Zurich, Switzerland, are: Werner Almesberger, Patrick Droz, Marco Heddes, Jean-Yves Le Boudec, Tony Przygienda, Linh Truong, and Colin West. Thanks to Gary D.

Figure 4 Widest-path routing example



Schultz for editorial assistance in the preparation of this paper.

*Trademark or registered trademark of International Business Machines Corporation.

Cited references

1. J.-Y. Le Boudec, "The Asynchronous Transfer Mode: A Tutorial," *Computer Networks and ISDN Systems* **24**, No. 4, 279–309 (May 15, 1992).
2. LAN Emulation SWG Drafting Group, *LAN Emulation Over ATM Specification—Version 1.0*, The ATM Forum, Worldwide Headquarters, 480 San Antonio Road, Suite 100, Mountain View, CA 94040-1219 (January 1995).
3. M. Laubach, *Classical IP and ARP over ATM*, RFC 1577, Internet Document (January 1994).
4. R. Hinden, *Internet Protocol Version 6 Specification*, Internet Draft (October 1, 1994).
5. The ATM Forum, *UNI Specification, Version 3.0*, Prentice-Hall, Inc., Englewood Cliffs, NJ, ISBN 0-13-225863-3 (September 1993).
6. The ATM Forum, *Interim Inter-switch Signaling Proto-*

col (IISP), Draft 94-0924R2, Worldwide Headquarters, 480 San Antonio Road, Suite 100, Mountain View, CA 94040-1219 (November 1994).
7. The ATM Forum, *P-NNI Draft Specification*, Draft 94-0471R3, Worldwide Headquarters, 480 San Antonio Road, Suite 100, Mountain View, CA 94040-1219 (1994).
8. A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley Publishing Co., Reading, MA (1975).
9. K. McCloghrie and M. Rose, *Management Information Base for Network Management of TCP/IP-based Internets: MIB-II*, RFC 1213, Internet Document (March 1991).
10. M. Ahmed and K. Tesink, *Definitions of Managed Objects for ATM Management Version 8.0 Using SMIv2*, RFC 1695, Internet Document (August 1994).
11. F. Baker and R. Coltun, *OSPF Version 2 Management Information Base*, RFC 1253, Internet Document (August 1991).
12. W. Denzel, A. Engbersen, and I. Iliadis, "A Flexible Shared-Buffer Switch for ATM at Gb/s Rates," *Computer Networks and ISDN Systems* **27**, No. 4, 611–624 (January 1995).
13. J. Moy, *OSPF Version 2*, RFC 1247, Internet Document (July 1991).
14. J.-Y. Le Boudec, R. Sultan, and B. Przygienda, *Routing Metric for Connections with Reserved Bandwidth*, European Fiber Optical Communications and Networks, Heidelberg, Germany (June 1994).

Accepted for publication April 12, 1995.

Robert A. Sultan *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 704, Yorktown Heights, New York 10598 (electronic mail: sultan@watson.ibm.com).* Mr. Sultan joined IBM in 1979, working on the design and implementation of manufacturing systems. He moved to the Computing Systems Department of the Research Center in 1981 where he worked in the area of telecommunications. In 1983 he joined the Network Architecture and Protocols Group to participate in System/36 APPN design and implementation as well as APPN performance studies. He was lead designer for a portable APPN implementation produced by the Experimental Systems Development Center. On assignment at the IBM Zurich Research Laboratory for four years, Mr. Sultan was project leader for the IBM 8260 Hub ATM Control Plane. He returned to the T. J. Watson Research Center in 1995, where he is currently manager of broadband networking software for the Advanced Networking Laboratory. Mr. Sultan received a B.S. degree from the Massachusetts Institute of Technology in 1968 and an M.S. degree in computer science from the Pennsylvania State University in 1979.

Claude Basso *C.E.R. IBM France, Le Plan du Bois, 06610 La Gaude, France.* Mr. Basso joined IBM in 1985, working on the design and implementation of call logging systems for telephone companies. He moved to the Communication Controllers Department in 1987 where he worked on the architecture and design of the IBM 3746-900, and then as manager of a software development group. In 1991, he joined the then-new Multiprotocol Hub department where he participated in the selection of the IBM partner and later in the elaboration of the longer-term strategy. A member of the system design group, Mr. Basso was actively involved in the definition of the technologies and products required to introduce ATM technology in local area networks. He was project leader for the IBM 8260 Hub ATM software (in cooperation with R. A. Sultan for the ATM Control Plane). He received a computer science engineer degree from the Ecole Nationale Superieure d'Informatique and de Mathematiques Appliquees de Grenoble–France (ENSIMAG) in 1981.

Reprint Order No. G321-5573.