
The Technical Challenges of High-Speed Switching I

3

ATM is not a bigger, brighter light bulb. Its benefits do not include simplicity. It will force one to do more planning and to understand one's application (and network) in greater depth than predecessor smart multiplexer or virtual private network solutions. However, first one must understand the basic problems endemic to high-speed switching and networks, the design compromises, and particular pathologies.

3.1 INTEROPERABILITY WITH LEGACY SYSTEMS

Legacy systems is a slightly pejorative term for the computer and communications systems one is currently using. The clear implication is that they await replacement by something better. Like money left one by a distant relative, there is also the implication that these systems were not one's direct doings. Less metaphorically, legacy systems were designed differently.

Unlike ATM, they generally used different technologies for the local area and wide area. WANs used leased circuits and SNA or DECnet or public networks (or public network technologies) and ran TCP/IP or X.25. LANs used 3270 or asynchronous terminal protocols or Ethernets or token rings running Novell's Netware or TCP/IP. Except for the significant exception of TCP/IP, the legacy of legacy systems was different technical approaches to the local and wide area and, inadvertently, extra complexity.

Another design difference is that of connection-oriented and connectionless. Current telephone and video distribution systems are connection-oriented. Some data networks like X.25 and SNA are connection-oriented. But the very numerous IEEE 802.3/5 LANs that interconnect with TCP/IP and Novell's Internetwork Packet Exchange (IPX)

are connectionless and have necessitated a complex accommodation. At this writing it is too soon to evaluate whether this complex bridging of network types will be successful.

Meanwhile, in the marketplace, where the time-honored way to market share and niche dominance is being the “first with the mostest,” what does one do when a standard is clearly needed and yet the standard is not there? The answer is the “prestandard,” or the proprietary implementation that makes the product useful to users and represents the company’s best guess at what the standard implementation will look like. As a consequence, most of today’s working ATM networks employ “prestandard” software. To their credit, realizing the importance of interoperability, few have tried to carve out proprietary niches; almost all have rushed to implement the new standards once blessed by the ATM Forum.

In the WAN area, ATM switches are likely to eventually replace smart multiplexers, packet switches, routers, and frame relay equipment. Depending on the context, ATM can provide more efficient multiplexing, better QoS guarantees, or simply speed. The anticipation is that the data communications protocols transplanted (SNA, DECnet, TCP/IP, IPX) will be encapsulated using AAL 5.

LANs can be viewed as a special case of legacy protocols. First, their ubiquity makes a smooth ATM transition an absolute necessity. Second, Ethernet, the most popular, is so unlike ATM that its efficient accommodation poses a significant technical challenge. Third, as noted earlier, providing higher performance LAN services has been the *raison d’être* of the early ATM switches.

With multitudes connected to Ethernet or token ring LANs, congestion at servers or network gateways often becomes severe. As a consequence, many organizations have moved to what are, in effect, two-level hierarchical LANs, often called “collapsed backbones.” These feature Ethernet or token ring work group LANs, often with local servers, that are then attached to departmental LANs for communications between departments, whether local or distant. Oftentimes the higher level LANs, reflecting the volumes of pooled traffic, use faster media, such as 16 Mbps token ring or fiber/copper distributed data interface (F/CDDI). ATM LAN switches, with their much higher speeds, are a natural in this role.

With the ATM Forum’s LAN emulation, one gets a marriage of necessity between two very different partners. The LANs user shares the bandwidth with neighbors, communicates locally via media access control (MAC) addresses, and employs variable-length frames to package data. In the case of Ethernet, all messages are broadcast to all users on the LAN segment. The ATM user, by contrast, sets up an end-to-end SVC connection prior to transmission, employs ATM addressing, packages data in uniform 53-octet cells, and tears down the connection after the

transmission is concluded. LAN emulation, often referred to as “LANE” (Phase 1), is the ATM Forum’s quick solution to getting these two very dissimilar technologies to work together. (A more general solution, incorporating both ISO layers two and three, is in development and is called “MPOA,” for multiple protocols over ATM.)

The LANE specification defines a standard way to correlate LAN MAC addresses with ATM virtual circuits. The LANE technology requires two specialized servers, the broadcast and unknown server (BUS) and the LAN emulation server (LES). The BUS and LES are typically implemented in software on a switch that connects the Ethernet ports to ATM.

The BUS and LES collect and store addressing information. When a workstation or server directly attached to the ATM network—called a LANE “client”—needs to contact a node on the legacy LAN, it issues a request that is received by the LES. If the LES knows the address, it responds. If not, it signals the BUS to broadcast the query over all the ATM virtual circuits to all nodes on the network. When located, the node responds with its MAC address and the BUS forwards it and the associated virtual circuit to the LES. The LES, in turn, forwards the information to the LEC and also updates its tables, allowing it to respond directly and avoid unnecessary future broadcasts. And to improve efficiency further, the LECs cache the addressing information so they do not make unnecessary LES requests.

Several other aspects about the LANE specification need be noted. First, the switch performing the LANE function is also a LEC, so when an Ethernet node tries to reach another station across an ATM backbone, the switch figures out which virtual circuit to use. Second, LANE defines a way for LECs to set up virtual circuits to transmit data. Known as data direct virtual circuits, they employ UNI 3.0 or 3.1 standards. Third, LANE also describes control VCCs, which are used to transfer cell control information between the LEC and LES or BUS. Finally, LANE specifies the segmentation and assembly (SAR) functions that render Ethernet or token ring frames into ATM cells.

Overall, probably the most important single feature (or limitation) of LANE is that it represents a MAC layer bridging (ISO layer 2) solution, which means that a router is needed to connect emulated LANs. And this limitation portends the coming of MPOA.

Below is an illustration (Table 3.1), employing the ISO’s seven-layer typology, but expanding the layer 2 data link shows where LANE functions in an Ethernet-TCP/IP environment.

Having described how LANE works and its place in the protocol stack, a not impertinent question remains: Does it work as described in the LANE 1 standard and can anyone demonstrate it today? Kevin Tolly in *Data Communications* magazine investigated in the September 1995

Table 3.1
LANE Functionality Within an Ethernet TCP/IP Environment

<i>LAN</i>	<i>ATM</i>	<i>LAN</i>
7 Application		Application
6 O/S-specific		O/S-specific
5 API socket		API socket
4 TCP		TCP
3 IP		IP
2 Ethernet MAC	Ethernet MAC ATM LUNI ATM AAL ATM Layer	Ethernet MAC
1 Physical	Physical	Physical

Key: AAL: ATM adaptation layer; API: applications programming interface; IP: Internet protocol; LUNI: LANE user-to-network interface; MAC: media access control; O/S: operating system; TCP: transmission control protocol.

Source: Ross, T. L., "ATM APIs: The Missing Links," *Data Communications*, Sept. 1995, p. 120. © 1995 The McGraw-Hill Companies. All rights reserved.

issue and found the answer was mostly no. Only FORE could demonstrate it and they relied on a number of proprietary features of their simple protocol for ATM network signaling (SPANS) software to do it [1]. In a return engagement by *Data Communications* a year later, 15 vendors were solicited. Six accepted: Bay, FORE, Cisco, IBM, Newbridge, and 3Com. Bay and Newbridge used proprietary LANE software, and Newbridge offered prestandard MPOA. Cisco's catalyst 5000-Lightstream 1010 combination and 3Com's Cellplex 7000 emerged as the top performers [2].

3.2 MAKING A CONNECTION

As noted earlier, making an ATM connection is analogous to making a telephone call. The switching resources have to be in place (signified by a dial tone). The circuit resources have to be in place (signified by the ring-

ing). And the connection resources have to be available and consenting (signified by the other party answering).

With ATM, a single SVC may traverse a half-dozen switches, which, conceivably, could be furnished by a half-dozen different switch vendors. These switches, link by link, must exchange information regarding connection QoS (CBR, ABR, etc.) and, once data transfer is in process, information on potential congestion via a “rate-based” congestion algorithm. As the congestion algorithm reflects a 1995 decision by the ATM Forum, most vendors have not implemented it, much less tested it in a heterogeneous environment. Today’s commercial ATM networks provided by interexchange carriers (IECs) and competitive access providers (CAPs) (MFS Worldcom [ex-WiTel, ex-MFS Datanet], AT&T, British Telecom [ex-MCI], Sprint, Teleport) avoid these problems by using one, or at most two, switch vendors to provide what is in effect a subset of ATM services.

The actual back-and-forth communication process setting up an SVC is called connection admission control (CAC) and was depicted in the previous chapter in Figure 2.5. In this negotiation, the requesting node informs the network of the type of service required, the traffic parameters of the data flows and QoS for each direction.

The traffic parameters are as below:

- *Peak cell rate* (PCR): The maximum rate at which cells can be transmitted across a virtual circuit, specified in cells per second and defined by the interval between the transmission of the last bit of one cell and the first bit of the next.
- *Cell delay variation* (CDV) or *cell delay variation tolerance* (CDVT): Measures the allowable variance in delay between one cell and the next, expressed in fractions of a second. When emulating a circuit, CDV measurements allow the network to determine if cells are arriving too fast or too slow.
- *Sustainable cell rate* (SCR): Maximum throughput bursty traffic can achieve within a given virtual circuit without risking cell loss.
- *Burst tolerance* (BT). In the context of ATM connections supporting VBR services, it is the limit parameter of the generic cell rate algorithm (GCRA) whereby conformance to the traffic contract is defined. The GCRA, in turn, is defined by two parameters, the increment (I) and the limit (L).
- *Minimum cell rate* (MCR): An ABR service traffic descriptor, in cells per second, that is the rate at which the source is always allowed to send.

The result is that CAC, seemingly straightforward as depicted in Figure 2.5, becomes a complex, three-dimensional (Service × Traffic

Table 3.2
Service Classes and Applicable Parameters

Attribute	ATM Layer Service Categories					Parameter
	Constant Bit Rate (CBR)	Variable Bit Rate (VBR) Real-Time	VBR Nonreal- Time	Available Bit Rate	Unspecified Bit Rate	
Cell loss ratio (CLR)	Specified ¹	Specified ¹	Specified ¹	Specified ²	Unspecified	Quality of service (QoS)
Maximum cell transfer delay (CTD) and mean cell transfer delay (CDV)	CDV and max CTD	CDV and max CTD	Mean CTD only	Unspecified 6	Unspecified	QoS
Peak cell rate and cell delay variation tolerance	Specified	Specified	Specified	Specified ⁴	Specified ³	Traffic
Sustainable cell rate and burst tolerance ⁵	N/A	Specified	Specified	N/A	N/A	Traffic
Minimum cell rate	N/A	N/A	N/A	Specified	N/A	Traffic
Congestion control	No	No	No	Yes	No	

Source: Alles, A., *ATM Internetworking*, San Jose, CA: Cisco Systems, 1995, p. 55.

Notes:

- 1) For CBR and VBR the CLR may be unspecified for CLP = 1.
- 2) Minimized for sources that adjust cell flow in response to control information.
- 3) Not subject to call admission control and usage parameter control procedures and may use different value from Section 3.6.2.4 of the User Network Interface 3.1 specification.
- 4) Represents the maximum rate at which the source can send as controlled by the control information.
- 5) These parameters are either explicitly specified for permanent virtual circuits or switched virtual circuits as defined in Section 3.6.2.4.1 of the User Network Interface 3.1/3.0 specifications.
- 6) The objective of the service is that the network does not excessively delay the admitted cells. Requirement for explicit specification of the CTD and CDV is for further study.

Parameters \times QoS Parameters) metric. Table 3.2 below reduces this complexity to two dimensions; note that a number of combinations are still not specified.

3.3 MULTIPLEXING AND DEMULTIPLEXING

Everyone is familiar in nature with how several small streams of water combine to form a swift larger one, and, conversely, how a large stream will divide into many small tributaries. The practice of combination for efficient transit and decombination for efficient distribution has been a fixture of modern telecommunications for decades. Most are familiar with the voice network “mux” and “demux” hierarchy of the public switched voice network. The basic customer unit (“stream”) is the 64-Kbps (DS-0) time division multiplexed voice circuit. Narrowband data circuits were created by digital access controllers (DACs) that demuxed the 64-Kbps streams further, usually in various increments down to 4.8 Kbps. In the other direction, large numbers of 64-Kbps circuits were “muxed up” or “rolled” into larger rivers by “channel banks,” 24 64-Kbps circuits into standardized DS-1s and further, 672 into proprietary DS-3 formats.

The down sides of this muxing up and muxing down were that it involved considerable hardware and the hardware tended to make the process inflexible and greatly slowed circuit provisioning, often taking a calendar month or more to install a data line of a certain capacity. With increased use of software control, the provisioning cycle has been lessened considerably, particularly in urban areas where the vendors have most of their resources (and customers) concentrated. The broadband dream of bandwidth on demand is through sophisticated software control, smart ATM switches, and enormous residual capacity in the vendor’s fiber-optic-dominated transmission network, the customer—not the vendor—will be able to instantly provision what is required when a particular application wants it.

Economically, this vision also subsumes a fiber-optic version of “all roads leading to Rome.” Indeed, a number of smallish communities have made considerable investments in communications infrastructures (via direct carrier subsidies) in order to attract (or retain) large teleservice or telemarketing operations. Many communities view their location on the “information highway” similarly to having harbor, bus, rail, and jet service. Nevertheless, in the absence of local, state, or federal subsidies, such vendor facilities must be cost-based and the resources available (links and nodes) in any particular location are ultimately finite. As more and more of the telecommunications market is deregulated and cost-based, huge service area inequities are likely.

Uncertain at this writing (Summer 1996) are what will be the popular sizes of bandwidth corresponding to the old electrical hierarchy's DS-0/1/3. The optical hierarchy provides an immense spread:

- OC-1 (51.84 Mbps);
- OC-3 (155.52 Mbps);
- OC-12 (622.08 Mbps);
- OC-24 (1.2 Gbps);
- OC-48 (2.4 Gbps).

It is safe to assume that the small sizes will be the most popular, but the decisions made by the cable and satellite providers of video to the home will greatly influence the equipment market and accelerate the popularity of particular channel sizes, just as the standardization of DS-1 (also known as T-1) did for that market earlier.

Whatever the channel size, ATM can be viewed as the latest and most sophisticated of statistical multiplexers. "Statistical" multiplexing is when several streams are combined on a channel in such a manner that the aggregate of the streams (such as bursty data traffic) can temporarily exceed the channel-carrying capacity without data loss. The burstiness is handled by temporarily buffering (storing) the excess. This, of course, adds to the transmission delay. In severe cases of excess offered traffic, conventional statistical multiplexers quiesce one or more sources. ATM allows what one might call a hybrid solution. Like conventional multiplexers, it allows for a protected partition for high-priority traffic (CBR, VBR) while statistically multiplexing lower priority traffic (ABR, UBR). This is illustrated in Figure 3.1. Because of the high bandwidth/data volumes involved, ATM also incorporates the option of selectively discarding low priority traffic that threatens to overwhelm the switch, as well as quiescing or throttling down the sources.

Last, mention must be made of ATM service multiplexers. Service multiplexers combine and decombine different services (voice, computer data, video) into one or more high-bandwidth ATM streams (pictured in Figure 3.1). They can be viewed as a special case of an ATM edge switch; their conventional telco counterpart is the digital access controller (DAC) or channel bank. Although the service multiplexer role seems straightforward in theory, it is almost never so in practice, as digital service units (DSUs) and specialized processors for encryption/decryption, compression/decompression, silence suppression/restoration, format or protocol conversion, and so forth are almost always present as well.

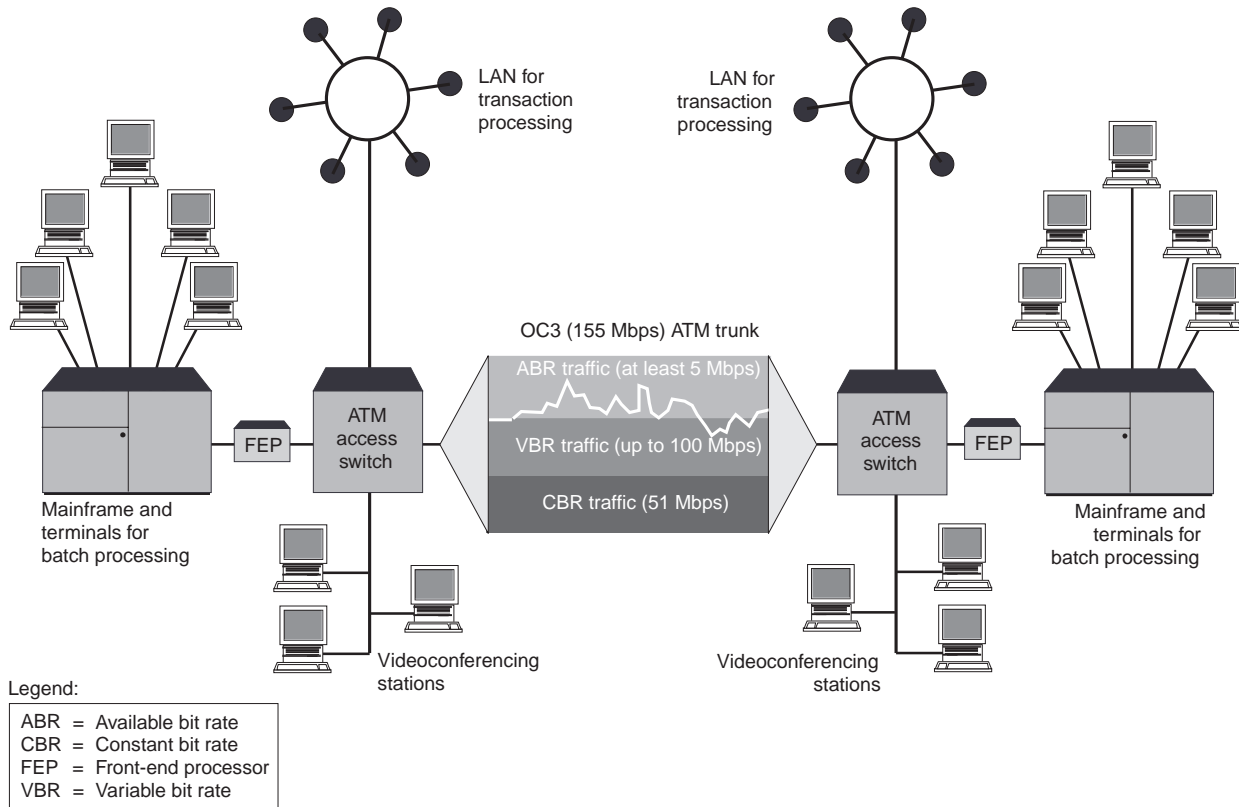


Figure 3.1 Statistical multiplexing. (Source: Hughes, D., and K. Hooshmand, "ABR Stretches ATM Network Resources," *Data Communications*, April 1995, p. 125. © 1995 The McGraw-Hill Companies. All rights reserved.)

3.4 ERROR CONTROL

In music, when one or more players flub their parts—make gross errors—the drill is to “take it from the top,” replay the piece from the beginning of a page of music. Similarly, present day error control usually consists of one or more retransmissions over the affected link. In data communications, the retransmission unit (frame, packet, cell, message) can vary as current layered protocols perform what amounts to redundant (or multi-layer) integrity checking. Nonetheless, most usually it is the data link layer (ISO 2) with its cyclical redundancy check (CRC) that occasions these retransmissions. Usually it is invisible to the user. Unfortunately, this is not always so and an egregiously noisy link can cause constant retransmissions and throughput degradation, called “throttledown.”

Note, too, that present day error control is usually based on a link or portion of an end-to-end connection. One is familiar with the party game where some nonsense phrase is successively whispered from one guest to another. After very few transmission over an error-prone link (guests whispering), the phrase, whatever its initial understandability, is reduced to gibberish. The fragility of the individual link relative to the integrity of the message is the rationale for link-by-link error checking.

In data communications one can well imagine the effects of pathological conditions on several portions of an end-to-end connection. The throttledowns become cumulative—and large. Various weird things can happen. When the end-to-ends are highly intelligent, such as two people on a transoceanic call, delays can be highly variable and the communication (technically, “user to user”) can proceed successfully. When a telephoner reaches an answering machine or voice mail service (technically, “user to application”) he or she can decide whether to leave a message or hang up. Other applications, such as automated teller machines (still another “ATM”), have a set limit to their programmed patience and will hang up. Most affected by throttledowns are computerized “application to application” conversations such as when a retail cash register autodial a credit or check approval service. In this case, the throttledown may cause a time-out, yet the cause of the (apparently) negative result may be ambiguous. For instance, a little old lady, invariably somebody’s mother buying a birthday present for a new grandchild, may be perceived a dead-beat and erroneously pinched.

As alluded to above, error control is one of today’s traditional data communications services that many believe is anachronistic. Critics point out that where DS-1 based networks promised error rates of 10⁻⁷, networks based on synchronous optical networks (SONET) advertise error rates in the 10⁻¹¹ range. These critics argue (1) that error control has no

more place on today's fiber networks than horses on superhighways, and furthermore (2) error control should be solely end-to-end, at the "edge of the network" rather than link-based.

More cautious commentators concede the truth of the lower error rate, but point out that with the increased data volumes (in the plumbing analogy, the increased diameter of the "pipes") the errors will occur just as frequently. They also note that dropped or switched bits may have out-sized effects on some applications, such as financial transfers or where the data is encrypted. In the short run—which may prove to be decades—industrial-strength error control (16- or 32-bit CRC) will remain a feature in high-speed data networks. Why? Because most of the data traffic passed over the broadband networks of the future will be encapsulated present day network protocols, and virtually all use 16/32 CRCs. As Ronald Reagan once rationalized his B-movie film career, "They didn't want art, they wanted it Thursday."

Having said that ATM will typically be transporting, via encapsulation, the error-control schemes of predecessor higher level protocols, how does ATM police its own act? ATM's error control is found in two general locations: the ATM header and as a component of ATM AALs. One octet of ATM's 5-octet header is dedicated to a header checksum. Error checking the AALs is summarized below:

- AAL 1: AAL 1 adds a 1-octet header to each data block; this header contains sequencing, timing, and error-correcting information. Single bit errors can be corrected if required.
- AAL 2: Neither the ITU-T nor ATM Forum have defined the specific format of AAL's segmentation and reassembly (SAR) protocol data unit (PDU). The best prognostication is that AAL 2, if eventually specified, will be similar to AAL 1 and will include sequence checks and will detect/correct single bit errors.
- AAL 3/4: AAL 3/4 adds a 2-octet SAR header (which contains a sequence number) to each block. This allows it to identify misinserted or lost packets. It also adds a 2-octet trailer that contains a length field (which allows it to count data field octets) and an error-correcting code to detect and correct single bit errors or detect multiple bit errors.
- AAL 5: AAL 5's SAR layer uses a single bit in the ATM header at the end of each SAR data unit. The AAL 5 convergence sublayer (CS) adds an 8-octet trailer in the last cell. This trailer has four fields that include a 2-octet length field and a 4-octet cyclical redundancy check (CRC). The CRC is capable of detecting multiple bit errors and misordered cells.

3.5 LATENCY

Latency is media delay. In cooking, there is latency between when one turns on the gas flame and when the fat sizzles in the fry pan. In high-speed network switching, with the potential at least for some links of relatively long segments of high-capacity fiber-optic media, dealing with latency entails provisioning large buffers. Although the large buffers induce a delay in themselves, they allow the dampening of its effects (technically, the delay variation). The buffer space required equals throughput x times round-trip delay y , the “delay bandwidth product.” In round numbers, this amounts to about 1 KB of buffer per kilometer, and about 1 GB of buffer per second [3]. Dealing with the maximum product necessitates a plan for buffer allocation (a series of independent buffers versus shared memory) and management strategies for handling congestion.

3.6 DATA PACKAGING

Packaging for data transmission (files, messages, frames, packets, cells, etc.) is directly analogous to the various containers that fill our spaces and litter our dumps. Similar to their corrugated and styrofoam-filled cousins, they vary by expense (ratio of data bits/packaging), layers of wrapping (protocol layers employed), security (encryption), tamper resistance (CRC), ease of handling (addressing), and reliability of delivery (retransmission).

One observation about packaging is that it has to be periodically rethought or reinvented. To take an example from the auto industry, in the beginning imported cars were treated as general cargo, stored in the holds, and winched out of the holds to dockside. Nowadays, the most efficient importers use custom-designed, drive on, drive off ships that look like floating, elongated parking ramps.

In today’s high-speed data transmission employing ATM, the data is typically encapsulated, preserving the current protocols, usually to preserve the addressing scheme of an extant application. (Managing this encapsulation process in ATM are the “adaptation layers” or AAL.) Thus one encounters use of the U.S. Department of Defense’s transport control protocol/Internet protocol (TCP/IP), the protocol equivalent of a humvee, being transported (to continue the metaphor) by the network equivalent of a French Train à Grande Vitesse (TGV). From an efficiency point of view, the ratio of overhead octets to data octets, it is not a pretty sight. On most applications where encapsulation is employed, the efficiency is certainly lower than that of 1960’s era asynchronous ASCII, where one wrapped a

7-bit code with a start bit, stop bit, and parity bit, achieving somewhat less than 70% efficiency, as some of the data were control instructions.

Prospects for new applications are much brighter. Faced with the requirement of delivering a steady stream of compressed video as economically as possible, and with error detection/correction a nonissue, ATM-delivered video will likely employ an AAL 2 standard optimized for the service.

Relative to a judgment on the efficacy of ATM's data packaging, several observations are germane. First, except in the case of the massive forward error-checking protocols used in some wireless applications, protocol overhead is the price one pays for applications flexibility and ATM is arguably the most flexible of current packaging schemes. Second, ATM package overhead was intended to be service-specific, with AAL 3/4 the highest and AAL 5 the lowest, but ultimately adjustable. For example, nothing precludes voice traffic from being carried VBR-RT or even ABR. Third, the ubiquity of broadband transmission makes bit and octet counting arguably less significant in the equation than formerly. Fourth, the lower level (ISO 2/3) protocol inefficiencies are often more than offset by presentation (ISO 6) efficiencies, particularly with compressed video where compressions of four and three to one are achievable without discernible impact on the application.

3.7 PERFORMANCE

High performance stems from fast components working together. The drop-nosed orange French TGVs not only have powerful turbine engines but computerized speed and suspension systems, and are supported by advanced switching systems, precise track alignment, and high-quality roadbeds. In today's electronics milieu, silicon and memories and buses offer, at a given time, certain price/performance ranges. Very high performance is available only at very high cost. As consequence, one tries to create a product where high performance is achieved by a combination of marketplace components and clever design.

The most common of these design tricks is parallel processing. Most are familiar with the concept of a bucket brigade, where buckets are handed back and forth along a single line. Communications people would characterize the typical bucket brigade as a "full-duplex, serial channel." The process could be speeded up incrementally by faster workers (components) or by having the buckets passed in a single direction (half duplex). The throughput, however, would scale linearly if another 10 bucket brigades were added in parallel. This is the general idea with parallel

designs, near-arithmetic scaling with near-arithmetic cost increases. And this design is particularly apt for high-speed communications where there may be multiple logical flows (four channels video, six channels computer data) sharing one physical channel.

The fastidious, however, will have noted the caveat “near” regarding both performance increases and costs in regard to parallel designs. Parallelism, even the most clear, always requires overhead processing. In the case of cell throughput, most high-speed switches require initial sorting in order to assign logical flows to processing streams. Where a logical flow may be split between two processing streams, recombination and ordering may be required. In the case of cost, common switch components such as power supplies and card cages and buses pose capacity limits beyond which costs rise nonlinearly; real-time control software brings its own limitations. In sum, parallel design employing merchant electronics that linearly scales by cost/performance, and that can be neatly packaged in popular increments (usually the powers of 2), is everyone’s dream—and no one’s reality.

With ATM, high performance is probably the single driving feature, even more than its facility to handle multimedia, deliver QoS, or provide a single protocol architecture for LAN/MAN/WAN environments. This emphasis, particularly on the high-speed transfer of computer data, can be expected to continue. As noted later in the vendor tables (Tables 5.1 and 5.2), few vendors provide OC-24 (2+ Gbps) switch interfaces today and few Bell Operating Companies (BOCs) and interexchange carriers (IEXC) offer transmission services at that speed. But, in a telling example of the role performance plays, the recently issued (8/23/95) draft Request for Proposals (RFP) for Post-Federal Telecommunications Services 2000 (PF2K), which reflects a compendium of United States Government requirements, requires responsive vendors to be able to provide ATM services up to and including OC-48 over the continental United States and to specified foreign locations beginning in 1999.

3.8 SEQUENCING

Many have found themselves at a children’s Halloween or Christmas celebration that concluded with the tots lining up on stage, each holding a letter card, intended to spell out a greeting to the loyal parents, and hilariously mis-sequenced. In data communications, sequencing is one of those tasks that everyone agrees is necessary but not everyone agrees on who should do it. It would be most disconcerting were video or voice transmission presented to the viewer or listener out of sequence. Indeed, such “flashbacks” or “voice-overs” are common tricks of the film maker. Nor is

sequencing necessarily real-time. One is familiar with common commercial transactions such as credit or balance checks (usually interactive) prior to withdrawals (usually batched).

Most current commercial transport networks accept sequencing as part of their task. Sequencing requires serial numbers, at least occasional sorting, and often waiting for retransmissions. Network minimalists hold that sequencing, as well as other grubby network administrative work, is inappropriate for a network whose focus should be the speedy transmission of billions of bits. They view sequencing the way some view bicyclists on freeways. They maintain that these tasks would be more appropriately accomplished beyond the network endpoints. Further, they point out that most network protocol stacks are full of what they see as petty, efficiency-robbing redundancies. Often they are correct. Network sequencing, an ISO level-three function, is almost always repeated at level seven in application areas like financial services or messaging systems. In God's eye, such duplication is inefficient; pragmatically, however, today's network consumers expect a sequenced product and are willing to put up with, and pay for, a good deal of inefficiency in communications. Moreover, were it otherwise—should their data arrive unsequenced—some would have to redesign their applications.

ATM virtual circuits, unlike the Internet, operate logically in a first-in, first-out (FIFO) manner. Although physically there is both service multiplexing (voice, video, computer data) and VC multiplexing (computer data 1, computer data 2, ...), the data are still FIFO. In the case of discarded cells under ABR, where retransmissions are necessary to restore a broken sequence, resequencing is performed, but at the packet level by the transport protocol (TCP, TP4) and not at the cell level.

3.9 BUFFERS AND BUFFERING

Buffers are life's waiting rooms. Just as the patients (presumably, the inexpensive resource) sit about reading old magazines in a physician's office so that access to the physician (the expensive resource) can be optimized, so it goes in automata. Given their usual connection with delays, buffers are viewed as a necessary indulgence to a world filled with traffic surges, speed mismatches, requirements for error correction, and other pesky but real-life concerns.

High-speed networks, however, have opened a whole new set of questions regarding buffers. Advances in memory technology have relatively recently allowed (by historical standards) huge buffers for small cost. Network processors of today can economically offer buffers that exceed the main memory of mainframes in the recent past. And symmetrically,

with high-speed networks, much is expected of them. One author has presented it as follows:

“A single-source buffer calculation for an OC-3 (155 Mbps) ATM connection spanning 3,000 kilometers (about 1,850 miles) would be as follows:

Bandwidth = 353,208 (OC-3 delivers 353,208 cells per second)
Round-trip transmission time = 60 milliseconds (this is calculated by multiplying the round-trip connection distance—in this case, 6,000 km—by the speed at which light travels through a fiber optic cable, which is 5 km/ms.) $353,208 \times .06 = 21,192$ cells” [4]

And, at least in theory, all or a portion of these 21,912 cells might be received in error and have to be retransmitted. The conventional wisdom is to bemoan the size and cost of these buffers. Others have pointed out that relative to the cost of the transmission pipe—itsself a kind of buffer—huge modern buffers are the cheap solution to many of the pesky problems noted.

3.10 QUEUES

Queues are the civilized way of waiting. One encounters them everywhere. Most banks impose what is technically a “single queue” (line) and “multiserver” (several cashiers) arrangement. Small airlines, particularly at small airports, often offer a single queue and a single server. Large airlines at international airports generally present a multiqueue-multiserver arrangement.

Everyone at one time or another has encountered blockage, or a queue with (seemingly) infinite holding time. In the communist Poland of yesteryear, for instance, a long line at shop was called a “sklep w ogon,” or “shop with a tail,” humorously characterizing consumer behavior in the face of scarcity. Those familiar with third-world telecommunications environments have their pet stories, funny after the fact, of waiting for dial tone, and once dial tone was obtained, repeatedly dialing a rotary phone seeking to obtain an intercity trunk. Reaching the desired party often became anticlimactic.

In high-performance environments, a particular queuing problem called “head of the line” blockage is important to avoid. Technically, it is akin to “page faults” in virtual memory operating systems and the “fatal

embrace” in database systems; it involves a processing stoppage as a result of resource contention. For example, one is waiting in a bank queue and the teller cannot complete one’s transaction without the authorization of a bank officer, who is herself busy.

Although in the past these sorts of contention problems have been cleared by timers—analogue to a dedicated bank officer to clear difficult cases—the modern approach has been to proliferate paths. Aided by low-cost digital electronics, the incremental cost of parallel queues is small when compared to additional WAN diversity, or worse, the unavailability of a highly visible application or service.

Everyone hates queues, and, in a sense, broadband communications represent one more attempt at achieving the perception of “queuelessness”—that somehow we will dispose of a huge number of bits in a magic pipe and we shall get everything we transmit instantly. As with other schemes, ATM will disappoint. Almost any cost-based scheme will involve class-of-service strictures (discussed below) wherein low-use time, often late night or early morning, will be bargain priced. In this scheme, any discount traffic that arrives prematurely, like the too-early party guest, will be queued and/or back-pressured (quiesced). Given the potential volumes, the caveats above on buffer consumption also are relevant: a switch, to protect the integrity of its buffers, must be exceedingly nimble at either refusal (quiescing the source), buffer and quiesce, or (worse case) buffer/overflow/quiesce/retransmit.

As noted, ATM’s technical approach is to minimize unnecessary queuing delay. However, a permissive entry policy moves the problem downstream into the ATM nodes and links. As ATM congestion control consists of a number of elements and is quite complex, it will be over-viewed in the subsequent section.

3.11 CONGESTION CONTROL AND FLOW CONTROL

Everyone who has tried to make the 8 p.m. Friday night show of a new and popular movie knows about congestion. Too many customers overwhelm the ticket booth (the admission control function) and, perhaps, there are too many to fit into the theater (the transmission buffer function). Flow control in the movie industry usually means portable stands with wind-out nylon belts to enforce a single or multiple server queues for admission, often in conjunction with back-pressure and informal counting to limit physical occupancy to the maximum permitted by the fire marshal. Occasionally, large numbers of customers are turned away at the ticket booth. Depending on the ticket policy, sometimes people with tickets are not seated, are recycled to subsequent shows, or are refunded

their money (although a “half-duplex” illustration with QoS, congestion, and flow control in telecommunications is pretty much the same).

Congestion control is probably the most humbling of telecommunications subjects. Communications engineers are brought up on the sequence: study the behavior, understand the behavior, automate the behavior. That important aspects of the First-World infrastructure work reliably (gas, electricity, telephone) serve as a testament to the efficacy of this credo and its statistical underpinnings. Nonetheless, one has come to periodically expect “spikes” in mass behavior that regularly lay low portions of the telecommunications fabric. For instance, in 1994, the Ottawa radio station CHEZ-FM offered 53 pairs of Pink Floyd concert tickets free to callers. An estimated 300,000+ call attempts overwhelmed a specially equipped exchange, causing delayed dial tone and other problems in an area of over 100 miles [5]. The problem will not go away, nor is it limited to public networks; if anything, rapid increases in televised direct sales, electronic “town meetings,” and home shopping channels promise new excitement.

In theory, all events, even in surge conditions, should generate routine responses. If the long distance circuits do not have adequate capacity on the primary path, the excess demand should be routed onto secondary paths and the calls should go through. If the resources available for re-routing are temporarily exhausted, one should get a recording saying so. So, too, with local exchanges experiencing surges; they should provide a recording announcing that all circuits are temporarily busy and directing one to call later. This is what is supposed to happen.

Unfortunately, this is not what happens all the time. The 1981 AT&T divestiture made competitors out of cooperators, postdivestiture “startup” companies are now billion dollar corporations, and telecommunications use has greatly increased. New services (mobile phone, mobile data, pagers, call waiting, messaging) have appeared and others (fax, dial-up data) greatly increased. No one—not AT&T, the Pentagon, the Federal Communications Commission, the 50 state regulators—even pretends to effective control. Ultimately, only the market, where you cannot successfully bill for a service that does not work, provides a discipline to this happy chaos.

There several reasons why contemporary efforts at congestion control have become less tractable. Some believe that our centralized models for handling congestion control are simply inappropriate to a semianarchic, distributed universe. For example, a marketing company using brand x equipment contracts with a network supplier using y equipment to carry z amount of data traffic. The company experiences a stunning response to a new service and generates z (cubed) traffic. The software configuration profile controlling congestion between brand x equipment

and brand y equipment has either been improperly set or is otherwise inconsistent and z (cubed) overwhelms a portion of the carrier's network, in the process denying service to other customers.

This would seem a trivial problem were it not that there are dozens of switch types with millions of lines of control software. In the course of a year, it is likely that this software will be stress-tested at thousands of locations by millions of customers with billions of accesses. Bugs or inconsistencies will appear. Human beings estimate traffic and configure the switches. Despite highly detailed standards and rigorous control and testing, the process is prone to specification errors. Weird things cannot only happen but the resulting pathologies are often propagated (e.g., one switch passes incorrect information to neighboring switches). In sum, historically, perfect hardware/software configuration control and switch management was unattainable in a highly homogeneous and hierarchical environment with relatively few traffic types (prevestiture AT&T). It is now much more difficult with a heterogeneous, distributed environment with many more players, with many more traffic types, and much higher volumes.

Congestion and flow control are among the most complex issues with the ATM technology. The reasons for this complexity are readily evident. First, ATM purports to function as a *service* multiplexer, harmoniously combining voice, video, and computer data according to QoS guarantees. Second, ATM's services, ABR and UBR in particular, are statistically multiplexed, which implies input flows, however brief, exceeding channel capacity. Third, ATM is a high-speed cell-switching technology where delays of any kind (queues, buffers) are to be avoided at all costs. Fourth, although ATM switches can protect themselves by selective discarding of ATM cells, the requirements of many data types, as well as the additional complications introduced by encryption and/or compression, make this option very unattractive from a throughput standpoint. Fifth, ATM is a so-called "unification" technology and will appear in LAN, MAN, and WAN environments where the fielded equipment will have large differences in capacity, particularly in buffering.

This complexity was much reflected in disagreements within the ATM Forum. In the Forum's recent decision in favor of "rate-based" flow control (where all switches in the connection path agree on incremental throughput adjustments to ABR/UBR flows), the merits of a competing, credit-based proposal were strongly argued. (With "credit-based" flow control, all switches in the connection path agree on the traffic adjustments on the basis of available buffer capacity.)

It should be pointed out early in any discussion of ATM congestion and flow control that today's ATM switches do not employ standard approaches. Today's public ATM networks are based on homogeneous

switches using proprietary (often termed “prestandard”) techniques. ATM networks employing switches from several manufacturers often are forced to rely on PVC connections and network interface cards (NICs) that limit input flows; others simply overengineer their networks and sacrifice statistical multiplexing.

With the various warnings and disclaimers past, what are the range of traffic and congestion control functions available to ATM? Basically there are nine such mechanisms, though (as explained below) no ATM switch manufacturer makes use of all them; they are as follows:

1. Connection admission control (CAC);
2. Usage parameter control (UPC);
3. Selective cell discarding;
4. Traffic shaping;
5. Explicit forward congestion indication (EFCI);
6. Resource management using virtual paths;
7. Frame discard;
8. Generic flow control (GFC);
9. ABR flow control.

CACs are the actions taken at SVC establishment to determine whether a connection can be “progressed” or not. To use a transportation metaphor, it is akin to a truck driver having his passport, license, registration, insurance, cargo permits, transit visas, and so forth in order prior to departing on an international run.

UPC, also known as traffic policing, is directly analogous to state troopers enforcing an automobile speed limit. In the case of ATM, it means that the ATM switch must enforce a particular connection-based traffic contract. In practice, this means limiting the peak rate of a connection to that of the slowest link along the path; the automobile analogy might be the reduced speeds encountered with heavy traffic.

Selective cell discarding is just what it implies—the ATM switch throwing away premarked cells considered disposable. An approximate transportation example would be, in emergencies, the banning of automobiles from the central city except for police cars, taxi cabs, and those vehicles marked “For Official Use Only.” In the case of ATM, it consists of discarding cells whose cell loss priority (CLP) equals 1, so to protect, as long as possible, $CLP = 0$ flows.

Traffic shaping (or pacing) refers to a mechanism that alters the characteristics of a stream of cells on a connection so to better meet QoS objectives. An automobile analogy to traffic shaping is the stoplights that one finds on some freeway entry ramps; they regulate (delay) access with the objective of maintaining a minimally acceptable speed on the freeway it-

self. Traffic shaping in ATM is limited by being required to preserve the cell sequence integrity of the connection. In practice, shaping almost always increases delay, measured here as the mean cell transit delay (CTD).

How and where in the network shaping is implemented is network-specific. The ATM Forum's UNI 3.1 specifies a dual "leaky bucket" mechanism for UPC and shaping can be visualized as the "front-end" of policing (Figure 3.2).

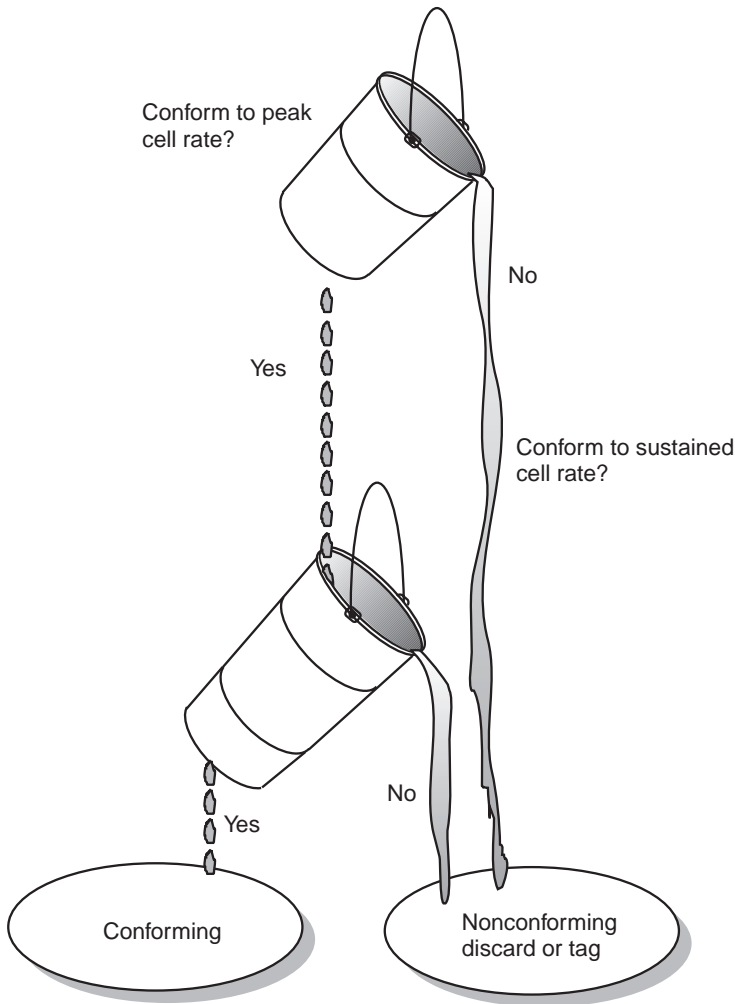


Figure 3.2 "Leaky bucket" mechanism to implement usage parameter control. (Source: Modarres, H., *Performance Considerations, Wide Area ATM Networks*, Herndon, VA: Newbridge Networks Corporation, 1995, p. 7.)

A “leaky bucket” mechanism implies a finite-sized bucket into which traffic flows and a maximum flow rate or “leak” from which traffic emerges. In ATM terms, the size of the bucket determines the cell delay variation tolerance (CDVT); the leak rate is the peak cell rate (PCR). Details of these algorithms are discussed at great length elsewhere [6,7]; suffice it to say that the effects of this electronic leaky bucket are to “shape” or “pace” traffic in the manner depicted on the left side of Figure 3.3.

Explicit forward congestion indication (EFCI) is a cell header setting that may be examined by the destination end system. Its highway analogy is the freeway signs that warn motorists of impending delays and possible alternative routes. The highway parallel is particularly apt because the ATM switch, like the tired motorist, may choose to do nothing. EFCI is optional for CBR, both kinds of VBR and UBR; only in ABR is the end system obligated to do anything upon receiving the EFCI.

Resource management using VPs allows one to implement a form of priority control by segregating groups of virtual connections according to service category. Essentially, it consists of intelligent planning or, more baldly, plain overengineering to avoid congestion. A highway parallel might be a toll road running adjacent to a free road where congestion could be avoided, literally, at a price.

Frame discard is based on the reality that if a congested network element needs to discard cells, it is oftentimes more effective to discard at the frame level than at the cell level. Often this shows up in vendor literature as the “early packet discard” feature. Where cell discard was likened to the emergency banning of private car traffic from a city center to avoid congestion, frame discard could be likened to the special banning of large trucks and buses during AM and PM rush hours.

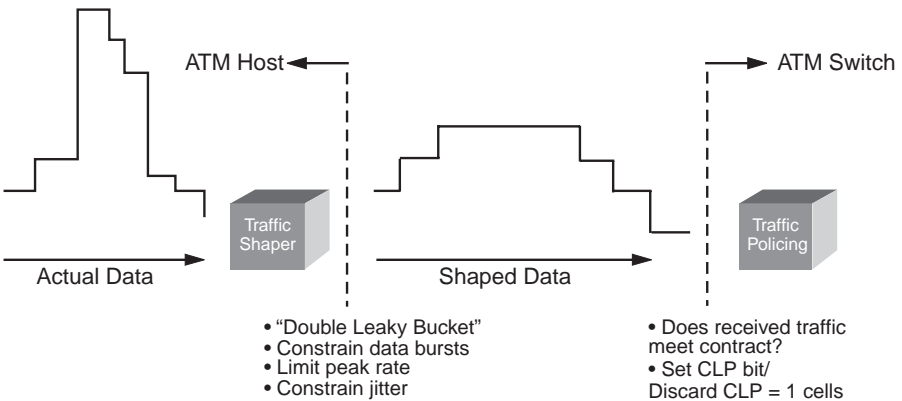


Figure 3.3 Traffic shaping and policing. (Source: Alles, A., *ATM Internetworking*, San Jose, CA: Cisco Systems, May, 1995, p. 56.)

Generic flow control (GFC) applies only to ABR service and is “not precluded” at the UNI for draft Traffic Management Specification 4.0. It is not expected to be implemented, but remains a tool in ATM’s conceptual toolbox for congestion control.

In ABR flow control, the source adapts its rate to changing network conditions. It can be likened to the ultimate “smart” highway where sensors and controls operate dynamically, raising and lowering the speed limits, advertising stoppages, and maximizing throughput. One can appreciate the complexity of this process. In the ATM Forum the debate between the credit- and rate-based constituencies was very vigorous, with the rate-based approach the apparent winner. The potential payoffs of ABR flow control are appropriately large: where all CBR and most VBR traffic is expected to use fixed bandwidth portions, ABR is likely to be home for most computer data traffic—and the source of statistical gain in the network. The mechanisms that make ATM’s ABR flow control possible—conveys state information about network elements to the source—are special control cells called resource management cells (RM-cells).

3.12 SUMMARY

As noted earlier, the technical challenges posed by high-speed networks are numerous—too numerous for a single chapter. In ATM, these attributes have been further disaggregated into a huge number of terms, each with its forbidding acronym (see Acronym List). At the extreme, they allow the ATM cognoscenti to converse in ATMTalk, a language combining the honorific and soporific, and often opening the door to pecuniary gain. But the concepts behind the attributes discussed in this chapter—interoperability, connections, muxing, error control, latency, packaging, performance, sequencing, buffers, queues, and congestion and flow control—are familiar. Most get a dose of them each morning driving to work.

End Notes

- [1] Tolly, K., “In Search of ATM LAN Emulation,” *Data Communications*, Sept. 1995, pp. 29–30.
- [2] Mandeville, R., and J.T. Johnson, “Forget the Forklift,” *Data Communications*, Sept. 1996, pp. 120–134.
- [3] Katevenis, M., Stefanos Sidiropoulos, and Costas Courcoubetis, “Weighed Round-Robin Cell Multiplexing in a General-Purpose ATM Switch Chip,” *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 8, Oct. 1991, pp. 1265–1279.

- [4] Hughes, D., and Hooshmand, "ABR Stretches ATM Network Resources," *Data Communications*, April 1995, p. 126.
- [5] Neumann, P. G., Moderator, "Risks to the Public in Computers and Related Systems," in ACM SIGSOFT, *Software Engineering Notes*, Vol. 15, Issue 70, (also available via <http://catless.ncl.ac.uk/Risks/15.70.html>).
- [6] Alles, A., *ATM Internetworking*, San Jose, CA: Cisco Systems, May 1995.
- [7] Partridge, C., *Gigabit Networking*, New York, NY: Addison-Wesley, 1994.