Asynchronous Transfer Mode (ATM) Switch Technology and Vendor Survey

Noémi Berry[1]

Report NAS-95-001 January 3, 1995

NAS Systems Division
NASA Ames Research Center
Mail Stop 258-6
Moffett Field, CA 94035-1000
noemi@nas.nasa.gov
http://www.nas.nasa.gov/~noemi
(415)604-4321

Abstract

ATM switch and software features are described and compared in order to make switch comparisons meaningful. An ATM switch's performance cannot be measured solely based on its claimed switching capacity; traffic management and congestion control are emerging as the determining factors in an ATM network's ultimate throughput. Non-switch ATM products and experiences with actual installations of ATM networks are described. A compilation of select vendor offerings as of October 1994 is provided in chart form.

---

1. Computer Sciences Corporation, NASA Contract NAS 2-12961, Moffett Field, CA 94035-1000

Table of Contents

1.0 Introduction

This paper surveys ATM switch vendors and technology by doing the following:

- Describe ATM switches and other products (Section 2, Products)
- Examine selected features of ATM switches (Section 3, Features)
- Experiences & comment on ATM (Section 4, Experience)
- Conclusion (Section 5, Conclusion)
- Vendors (Section 6, Vendors)
- References (Section 7, References)

This document is neither a tutorial nor in-depth description of ATM; rather, it is intended to be something of a buyer's guide. However, for the uninitiated reader, the following is a brief description of some of the salient points of ATM.

1.1 ATM description

Asynchronous Transfer Mode (ATM) is a switching technology which multiplexes and switches cells from multiple sources to multiple receivers. ATM cells are fixed-sized at 53 bytes, with 48 bytes of payload. The small fixed size allows efficient hardware implementations of switching fabrics. Unlike most packet switches, ATM switches are not store-and-forward switches, thereby reducing critical delays in the node. However, many ATM switches contain input and/or output buffers for traffic management purposes.

ATM is a connection-oriented protocol. Each connection has a Quality of Service (QoS) associated with it, negotiated during the startup of the connection. ATM will drop cells if necessary to meet its connections' QoS requirements, and the higher-lever protocols must recover from dropped cells. Elaborate traffic management schemes determine which cells must be dropped in order to maintain QoS to which the network has committed for all its connections. Congestion control schemes keep traffic flow smooth to prevent conditions in which cells would be dropped.

ATM cells are sent along Virtual Channels (VCs), which are transported within Virtual Paths (VPs). ATM connections are identified by a VPI/VCI (Virtual Path, Channel Identifier) pair. Connections are either Permanent Virtual Circuits (PVCs) or Switched Virtual Circuits (SVCs). SVCs use dynamic routing and load balancing, where PVCs are strictly static.

To encapsulate larger data units, an ATM Adaptation Layer (AAL) is used. A number of different AALs are used for segmentation and reassembly (SAR) functions, depending on specific needs of a connection. AAL 3/4 and AAL5 implement connectionless services and are of greatest interest for carrying IP over ATM.

ATM was intended to run over SONET carriers, with OC-3 (155 Mb/s) as the most common rate. Other physical-layer signalling is also commonly used, such as TAXI in the local area and DS-3 in the wide area. The signalling between a host and the network is known as a User-to-Network Interface (UNI). The signalling between two switches within the network is known as a Network-to-Network Interface (NNI).

Standards are developed by the ATM Forum, a group of 500 companies representing all sectors of the communications and computer industries, as well as a number of government agencies, research organizations and users.

See [1], [2] or [3] for more information on ATM.

2.0 ATM Products

ATM switches are only one (large) piece of the ATM picture. To deploy an ATM network in the LAN may require additional hardware and interfaces. Many vendors offer "solutions" for ATM access that do not require buying switches, that allow legacy LANs to interface with ATM equipment or services. ATM services can be purchased through service providers, though the customer needs CPE (Customer Premises Equipment) ATM equipment to connect to the service.

While this paper is intended to focus on switches, other ATM products deserve note as well. Not all ATM products fit into easy categories, but the one thing they have in common is that at one end or the other, or both, ATM cells come in, or go out. The following are ATM products described in this section.

- ATM switches
- ATM host adapters
- Routers with ATM interfaces
- Hubs with ATM interfaces
- ATM CSU/DSU
- LAN-to-ATM access devices
- Other

2.1 ATM switches

ATM switches may be loosely categorized into the following: Carrier, Enterprise/WAN and Campus/LAN. These categories are not rigid and some vendors' switches can span more than one category. Prices depend heavily on type and configuration.

Carrier switches provide something similar to a phone company service, in which the customer has no equipment on site, the switch can be made fully redundant (processor and fabric), the internal switching capacity is very high (up to 10Gb/s), and the cost of a switch is typically in excess of $100,000. Physically the switches are kept in a central office, are large rack-and-stack types, and require DC power. Carrier switches support WAN and SONET interfaces such as T1/T3, DS1/DS3, OC-N, generally over singlemode fiber and copper.

Enterprise networks are wide and metropolitan-area oriented, offer some redundancy and support interfaces to WAN and MAN signalling such as T1/E1, T3/E3, DS1/DS3, TAXI; with single mode fiber, coax, copper and some redundancy. An Enterprise switch is often used as a carrier edge node, the point of access to a carrier network.

Campus/local ATM switches are LAN oriented, have less switching capacity, processing power and ports; and are physically smaller and much less expensive than carrier switches (under $50,000). Campus switches provide interfaces to LANs such as Ethernet, token ring, FDDI and HiPPI; and DS3, TAXI and SONET interfaces over multimode fiber, coax, copper and UTP. Features such as LAN emulation (LAN switching over ATM) and virtual LANs are common. Many vendors who make campus ATM switches also make ATM interfaces for existing LAN products and hosts, such as routers and workstations.

In sum there are three basic types of ATM switches:

1) Carrier (CO, redundant, high switching capability, expensive)
2) Enterprise (WAN/MAN, some redundancy, single-mode fiber)
3) Campus/Local (cheaper, legacy LAN interfaces, multi-mode fiber)

This survey will emphasize Campus and Enterprise switches (some don't fit easily into a single category), though features of Carrier switches are worth noting. Only currently available switches are mentioned, though many more have been announced.

2.2 Host interfaces

To attach a host (workstation, printer, file server) directly to an ATM switch or other device, a host adapter card that plugs into the host is needed.

Host adapter (or host interface) cards perform a conversion of data units to ATM cells in the ATM Adaptation Layer (AAL), five of which have been accepted for consideration by the ITU-T (formerly CCITT). AAL 1 is for Constant Bit Rate (CBR) services (e.g. voice). AAL 2 is for Variable Bit Rate (VBR) services with required timing between source & destination (e.g. audio, video). AAL 3/4 is for VBR connectionless services. AAL 5 is a simpler version of AAL 3/4, with less error checking and protocol overhead. Adapter cards also perform the UNI signalling that set up a UNI to the ATM device.

Network Interface Cards (NIC) are the cards that provide ports on the switch. They do NNI signalling and do not contain AAL functionality or offer user connections.

All switch vendors make host interfaces, as well as many other companies such as Alantec, Texas Instruments, Sun and HP.

2.3 Routers with ATM interfaces

ATM switches' basic function is to switch ATM cells. Switches are connection-oriented and produce a flat topology. Routers discover addresses and network topology changes, implement subnetworks with hierarchical routing, perform media translation and protocol conversion, and provide dynamic rerouting.

Routers equipped with an ATM card can operate as an ATM access device by converting and routing cells to an ATM switch or service, or routing LAN traffic to a device that performs the AAL functionality (cell conversion) and then forward the cells to an ATM switch. An ATM module in a router can even function as a switch by switching ATM traffic between multiple ATM interfaces, with UNI functionality over the usual ATM interfaces. Cisco and Wellfleet both make ATM interfaces for their routers.

2.4 Hubs with ATM interfaces

Hubs are often used as the basic network building block to which to attach endstations and multiplex their LAN traffic to a single stream. They can be used in a hierarchy to higher-level hubs, sometimes via a higher-speed protocol, such as FDDI and eventually ATM. A hub hierarchy is often used to concentrate access among many individual users to a shared resource such as a server or router.

Hub ATM interface cards include AAL functionality and UNI signalling and permit hub communication across ATM. Companies making ATM interface cards for hubs are, not surprisingly, those already making hubs, such as 3Com, Wellfleet, Synoptics and Cisco.

2.5 ATM CSU/DSU

In addition to UNI and NNI, the ATM Forum has adopted the Data Exchange Interface (DXI), a method for low-speed ATM access allowing users to connect their legacy LANs to ATM services without paying for their own DS-3 or OC-3 line. ATM DXI specifies the interface between traditional network products (such as a router) to a box called the ATM CSU/DSU (Channel Service Unit/Data Service Unit) also known as a "cellifier," which provides the conversion to an ATM UNI to an ATM device. The DXI operates at low speeds, up to 50Mbps, and allows routers talk to packet-based interface instead of a cell-based interface[1]. Routers handle frames and packets but don't fragment them into cells; DSUs fragment frames and packets into cells and forward them over the UNI to

1. (supporting V.35, RS449 or HSSI)

the ATM switch or service. The ATM DXI accommodates AAL 3/4 and/or AAL 5. Companies that make DSU/CSUs are Network Systems and Cisco.

2.6 LAN-to-ATM access devices

To use ATM in existing LANs, LAN-to-ATM access devices convert legacy LAN traffic into ATM cells. These gateways allow a local ATM network to be used as a backbone transparently and often include LAN Emulation functionality. A UNI is set up between the LAN gateway and the ATM switch.

An example is Synoptics' Ethercell device, which multiplexes multiple Ethernet streams into one ATM stream to Synoptics' ATM switch. An advantage of such a device is that each Ethernet port can use Ethernet's full 10Mb/s bandwidth potential, taking advantage of ATM's nonshared media. Another example is Fore Systems' LAX-20 LAN Access device into which modules for Ethernet, FDDI and Token ring convert LAN traffic into ATM sent out on one or more ATM modules.

A LAN-to-ATM interface may also occur at a port on an ATM switch, with an integrated UNI. For example, and ATM switch may have token ring or frame relay ports that take the respective protocols' frames and convert them to ATM cells right there at the port. Lightstream is an example of vendor that provides "edge ports" for Ethernet, FDDI, token ring, and frame relay.

2.7 Other

Other ATM products are anything that produces ATM cells and/or sets up a UNI to an ATM switch. An example is Fore Systems' video adapter that converts video signals from a video device (camera, player) to ATM cells and sends them to Fore's ATM switch. Another type of product is ATM multiplexers/concentrators, which multiplex traffic from multiple ATM interfaces to a single stream.

ATM test tools and cell-sniffers are also coming out. HP has a useful but expensive (appx. $100,000) ATM cell tester (ES4210). Microwave Logic and Adtech are working on less expensive ATM testers.

2.8 Internetworking

Much ado is made about interoperability testing, but so far this has been mostly focused on testing host interfaces with various ATM switches, since the UNI standards are established. The UNI standard permits two switches to be interconnected as though a switch was a host, but this is without running an NNI across. Without an NNI standard or compatible prestandard NNI signalling, switches from more than one vendor cannot exist in the same network (except as mentioned above, as though a switch were a host, allowing only UNI signalling to be used). Eventually an SVC standard will permit SVCs across different vendors' switches. Despite standards, one vendor's ATM products will work best with that vendor's ATM switch, since the vendor will ensure interoperability between its own products.

3.0 Features

The features are divided into the categories outlined below, with descriptions following below.

Architecture/Resources
- Switch architecture
- Output/Input buffers
- Switch/call control
- Maximum number of ATM ports
- Number of VPI/VCIs

Performance
- Switch capacity
- Switch transit delay
- Connection setup time

Interfaces/Compliance
- ATM Interfaces
- Non-ATM interfaces
- UNI signalling

Software/Features
- Multicast
- SVCs
- NNIs
- Traffic management
  - Traffic policing
  - Congestion control
- LAN Emulation
- Network Management
- Extras

Cost of typical switch configurations
- 16-port LAN switch
- 64-port Enterprise switch
- 128-port carrier switch

Architecture/Resources

3.1 Switch architecture

A switch's basic function is routing cells. ATM's small, fixed-sized cells allow efficient hardware design of the switch fabrics that perform the switching and routing. As such, the switch fabric design is the core of the switch.

The basic models of ATM switch fabric architectures on the market are 1) Time division multiplexed (TDM) bus and 2) Space division switched matrix. The major architectures in use are derived from one or both of these models.

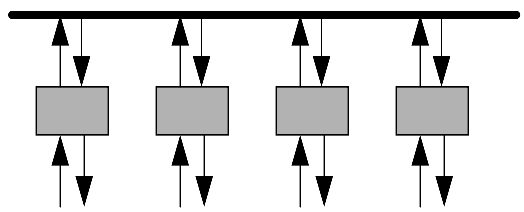
In a **TDM single bus** switch, the bus transfers cells from a source network interface (port) to one or more destination network interfaces (ports). Output buffering of cells may be provided by queues between the bus and ports, and by queues at each input or output port.

As long as the capacity of central switching fabric (bus) exceeds that of the individual ports, there is no contention at the ports or at the bus, and cells pass through the switch fabric with no delay (however, there may be delay at the output buffers). This architecture is called "non-blocking," since there is no port contention, provided the bus speed exceeds aggregate port speed -- the performance bottleneck is the bus.

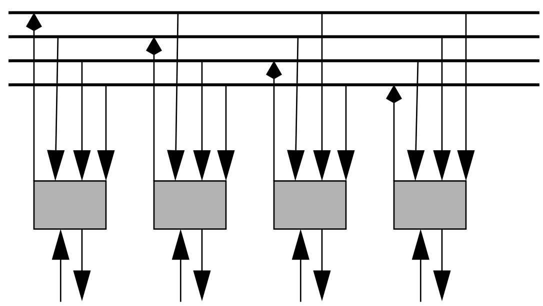
A TDM bus inherently multiplexes multirate traffic by simply using less time slots for the slower traffic. Cell replication is easily supported with no extra hardware complexity by transporting a single cell over the bus, and each port that is part of a multicast reads the cell.

The chief disadvantage of a single bus architecture is that it does not scale up well for adding ports, since more ports will quickly exceed bus capacity, introducing port contention and exceeding buffers, reducing performance. Its advantages are that it is simple and inexpensive for a small to medium number of ports and is contentionless (given that aggregate port capacity does not exceed bus capacity).

A **broadcast matrix** (also referred to as a **broadcast bus**) is a multiple-bus architecture that employs aspects of time-division and space-division switching. It is based on the "Knockout" architecture, and can appear in a crosspoint configuration. A broadcast matrix uses multiple (time-division) busses to separate (space-division) its traffic. Since it has no switching elements, its fabric is non-blocking due to no points of contention. There is one bus per port, with each of N ports listening to N-1 busses and broadcasting to the Nth. With every added port, there is added bandwidth from the added bus, and this architecture requires no extra complexity for cell replication. Though the fabric itself is non-blocking, a bus arbitration scheme and input buffering is necessary to prevent flooding an output port. The complexity results in a costly switch that is mostly used by carriers. Many vendors advertise the fabric as a "busless" matrix, meaning it does not employ a single TDM bus, though a broadcast matrix is technically a multiple-bus architecture.



4-port TDM single bus



4-port broadcast bus

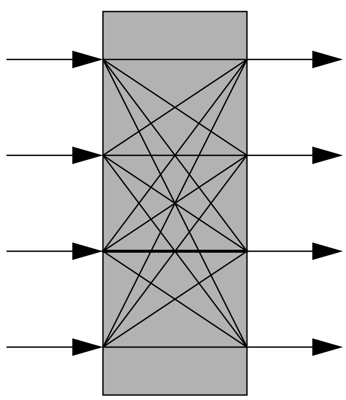
Space division switches typically have singlestage fabric or multistage matrix architectures. Matrix architectures interconnect multiple copies of a basic building block called a *switch element*. A switch element is a simple circuit that routes a cell from an input to an output, possibly buffering, and that occurs at the intersection of two or more "wires" in the fabric.

In a **singlestage fabric** architecture, each input port is direct-wired with a unique path to each output port (one giant switch element, if you will). This does not scale well for increasing ports, and some vendors offer more ports by tying together their basic fabric into multiple stages.

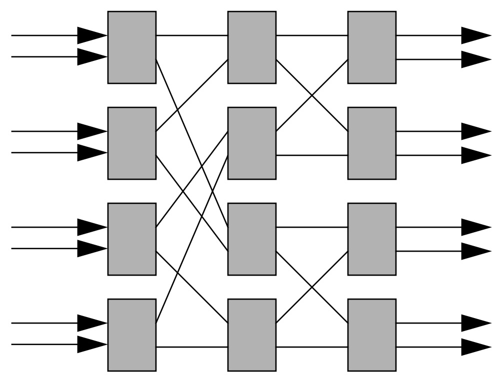
Usually singlestage fabrics appear in a crosspoint configuration. Crosspoint (or crossbar) architectures use an N x N port design, though this can refer to N slots, with slot cards having their own switching capability and containing one or more ports. An N x N crossbar switch may be able to switch N full-duplex connections (a la HiPPI), or may contain N slots with 4-port cards available for a total of N * 4 ATM ports. (Some classic crosspoint switching designs use a square matrix with N^2 internal crosspoints, providing blocking and re-routing opportunities in the switch, but typically crosspoint in the ATM world refers to contentionless single-stage fabrics).

In a **multistage matrix**, cells *self-route* through internal networks of switch elements, exiting the switch element on the best path through the network to its ultimate output port. Switch elements can have two or more input and output ports, and can be tied together in one of several designs for multiple stages[1].

Cells pass through the fabric synchronously so that all cells enter a stage on the same clock cycle. Depending on the internal network architecture, more than one cell may arrive at a switch element in one cycle. To avoid cell-dropping, switch elements contain internal buffers capable of storing several cells and resolving contention. Output buffers usually exist at the switch output ports.



4-port single-stage fabric



8-port multistage fabric (3 stages)

The diagrams show N inputs and N outputs, for effectively N bidirectional ports, without actually showing the wiring to a single physical port.

1. The number of stages can be computed as follows: $S = \log_M N$, where S is the number of stages, M is the number of input/output ports to a switch element, and N is the total number of input ports to the switch.

For example, the 16-port Synoptics switch has 2-port switch elements arranged in a delta network, so N = 16, M = 2, and therefore number of stages S = 4. A single-stage 12x12 matrix has N = 12, M = 12 and therefore stages S = 1. Networks of switch elements can be arranged in delta, Benes, Banyan or Clos designs for multistage matrices.

Self-routing matrix architectures do not support cell replication as inherently as those based on a bus. One strategy is to “fan” a cell from an input port, duplicating it as it passes through each stage to the output ports involved in the multicast. Some architectures do this in a separate copy network to reduce congestion in the routing network; others reduce delay by performing the cell replication and routing in the same network.

The multistage matrix concept is scalable to support more ports by adding stages. However, additional stages increases transit delay as a cell must pass through more stages, as well as increased opportunity for contention and delay at the switch elements in the extra stages.

The chief disadvantage of multistage matrix architectures is the possibility of buffering and delay at a switch element or port at any stage, since more than one cell can arrive at a switch element at a time. In a lightly loaded network, a cell can conceivably reach its destination with no buffering or contention, but this design is more sensitive to traffic characteristics. With the assumption of uniform, non-bursty traffic, in which the load is spread out evenly across the fabric, space division switches perform well. This is still true as more stages are added to increase the number of ports, so in that respect the matrix scales up well for more ports.

However, ATM traffic characteristics vary and may be non-uniform and bursty, producing bursts of traffic in concentrated areas and increasing the likelihood of “hot-spots” (areas of high congestion). With fewer ports across which to spread traffic, a matrix may not scale *down* well to the case of fewer ports (such as in a typical 16-port Campus configuration).

Space division switching has a well-known theoretical throughput upper bound of 58%. This result came from analyzing input vs. output queuing, and with an assumption of a uniform distribution of traffic load, which may not be realistic in ATM networks. The main reason for the 0.58 throughput is the head-of-line blocking (contention) in input queues. A blocked cell at the head of the queue can block cells behind it in the queue whose paths up ahead are not blocked. However, with only output queueing, throughput can be much higher since there is no head-of-line blocking. Needless to say, the benefits and liabilities of space-division switching is an ongoing research topic.

<u>Architecture issues</u>

A key issue in architecture design is **blocking probability.** *Non-blocking* architectures guarantee an absence of internal conflicts. In *blocking* architectures, paths from input ports to output ports share links between stages. *Virtually non-blocking* indicates a very small blocking probability in a blocking architecture.

All switches must employ buffer management and traffic management schemes to compensate for potential cell loss, so a blocking switch will not automatically perform worse than a non-blocking switch based solely on the blocking probability of the architecture. Any fabric can overwhelm a single output port. Some vendors whose aggregate port capacity can well exceed the fabric capacity maintain their switches are non-blocking, through the use of a clever buffering scheme. This guarantees that cells can pass through the fabric at the claimed fabric speed, though does not always directly translate into throughput.

Blocking performance is sensitive to switch architecture and traffic characteristics. As discussed above, a TDM bus can be non-blocking provided the aggregate port capacity does not exceed the bus capacity. Space division switch blocking probability depends on the number of paths between switch elements. The more paths, the lower the blocking probability. If there are as many paths between switch elements as there are input ports, then it is non-blocking, though this is costly. Most space-division switches, particularly the campus and enterprise type, are virtually non-blocking.

Another important issue in ATM switch technology is **scalability**. "Scalability" as advertised by vendors can refer to 1) increasing the number of ports 2) increasing the speed of the port interfaces 3) increasing distances between switches; all with no major architectural or software changes nor significant performance loss. A vendor may say their architecture is "scalable" but scalable in which aspect will vary. (ATM itself is intended to be geographic distance scalable; however this is not strictly a fabric architecture issue.)

Most users will be primarily concerned with performance and reliability, but issues in scaling and response to traffic characteristics deserve consideration before committing to a particular architecture.

Fabric Architecture comparison

	SS Fabric	MS Matrix	Brcast matrix	TDM Bus
Fabric blocking probability	Zero	Medium	Zero	Zero
Cell replication complexity	Fair	High	Low	Low
Scalability (fabric speed)	Good	Fair	Fair	Poor
Scalability (number of ports)	Poor	Good	Fair	Poor
Buffering in fabric	Likely	Required	Likely	Unlikely
Input buffering	No	Possible	Possible	No
Cost to produce	Low	Medium	High	Low

SS = Single-Stage
MS = Multi-Stage
Brcast = Broadcast

3.2 Buffers

A switch's buffering strategy plays at least as important role in a switch's throughput as the type of switching fabric employed, since it is the output ports at which the risk of cell loss is highest.

The tradeoff in buffer design is latency vs. throughput. Buffers are needed to hold cells in case an input or output port is busy, lest the cells be dropped. Holding cells in a buffer can increase delay, but not buffering cells (i.e. dropping) can decrease throughput. The smaller the requirements for cell delay variation (jitter) and cell delay, the smaller the buffers should be. The smaller the requirements for cell loss, the larger the buffers should be.

Switches can contain internal and external buffers. Internal buffering is usually employed between switch elements of a matrix architecture. External buffers occur before or after cells pass through the switch fabric, and hence can exist at input or output ports. The most common buffers seen are output (external) buffers. Internal buffers usually occur as small buffers between switch elements in multistage matrix architectures.

Buffers are a convenient place to employ traffic management schemes to examine and select cells for tagging or dropping. Also, buffers can be arranged in multiple priority queues according to traffic type. Many vendors implement multiple priority queues in their output buffers. Output buffers can be per port, per group of ports, or one large buffer allocated per connection.

Buffers can be organized as input, output, shared input, shared output, or shared input-output queues. Input queueing is easiest to implement, but can severely degrade performance due to Head-Of-Line (HOL) blocking, where a blocked cell at the head of a queue prevents cells with unblocked paths from proceeding. However, the HOL problem can be alleviated by mixing with a carefully chosen output queueing strategy, or with other operations such as windowing [16], and so does appear on some switches. Output queueing has been shown to be theoretically optimal, with shared output queueing providing the optimal delay-throughput performance. On a port or connection basis, output buffers can be drawn from a common shared buffer pool (AT&T), though some vendors take the opposite strategy and use non-shared output buffers (Stratacom) so as to guarantee buffer availability to every port or connection.

Most switches use output buffers, the majority of which are shared. Input buffers are more likely to appear on larger switches, particularly ones which have direct interfaces to other protocols (e.g. a native Frame Relay or SMDS interface).

With more ATM networks in use, the impact of buffering, queueing and traffic management strategies on throughput is rapidly emerging as a major issue.

3.3 Switch/call control

One of ATM's main function is setting up calls, or Virtual Channel Connections (VCCs). Call control determines the route of the call, grants resources according to desired quality of service, assigns VPI/VCI, and establishes the connection.

In a distributed switch control architecture, each switch has its own CPU and software with which to perform call control tasks. With a centralized switch control, a single computer performs control tasks. Current central control systems, usually a workstation, are connected to the ATM LAN by Ethernet; true ATM connections can be expected soon.

The advantages of a centralized switch control system are that it can use a large, capable machine for call processing, lower cost for processing power, and central administration for features such as virtual LANS. It is also easy to implement, maintain and upgrade. The disadvantages are that the central controller and the switch to which it is attached are both single points of failure, and the workstation is a bottleneck for call processing. It does not scale as the network grows and must have its capacity increased as switches, links and hosts are added to the network. Recovery time from a network outage is an issue since all switches must reach the controller during recovery. Therefore, centralized call control systems are only practical in LAN environments.

Centralized systems usually offer redundancy by adding one or more additional controllers on the network. If the call control system or the switch to which it is attached fails, then a redundant controller takes over. If the network becomes segmented, then the controllers manage the sections of the network they can reach. The pitfalls of redundant controllers are synchronizing redundant controllers' information (a basic problem in distributed systems), and it is still possible to segment a network without a controller that can reach it. Presumably the topology will be designed to avoid this situation, but the controller locations become an extra factor in topology design.

The advantages of a distributed system are robustness, quick connection setup and fault recovery, and inherent scalability for increased distances and network size (density, number of hops between endpoints). In a distributed system, in order for the entire network to be unreachable, every node has to be down. Call processing power is increased with every switch added to the network. The chief disadvantage is the increased cost of a switch for the on-board processor, in addition to added complexity for software upgrades and backward compatibility with old switches in a network.

Most vendors use a distributed system, primarily for reasons of robustness. Some say each switch has its own processor, but certain functions are relegated to an attached workstation, blurring the distinction between central and distributed.

Call Control Comparison

	Central	Distributed
Robustness	*D*	*A*
Distance scalability	*D*	*A*
Network size scalability	*D*	*A*
Connection setup speed	*D*	*A*
Fault recovery speed	*D*	*A*
Information exchange complexity	*A*	*D*
Cost per switch	*A*	*D*

A = Advantage
D = Disadvantage

3.4 Maximum number of ATM ports

Most campus-type switches offer at least 16 ATM ports, and most intend to offer up to 64. So far none of the campus switches have released a 64 OC-3 port switch, though several (Fore, SynOptics) intend to. 64 OC-3 ports feeds in almost 10 Gbps, of which currently only carrier switches are capable of carrying in their fabrics.

The more ports a switch has, the flatter and more interconnected the network topology can be. More ports also facilitates redundant links for critical connections. The cost of the extra ports may not be justified for smaller LANs that don't need as many as 64 per switch.

For the smaller-capacity campus-type switches, scaling up number of ports presents internal architectural problems. Single-bus architectures lose their non-blocking advantage unless the bus can be sped up to meet the aggregate port demand, which is prohibitively costly. Matrix architectures must add more stages, which works well but increases points of contention and increases the number of stages through which a cell must pass, increasing transit delay.

3.5 Number of VPI/VCIs

ATM connections within a switch are identified by a combination of a Virtual Path (VP) ID (VPI) and a Virtual Channel (VC) ID (VCI). (The VPI/VCI combination is often referred to as VPCI since there is some variation in how vendors allocate these IDs.) VCs are transported within VPs, which may aggregate VCs together or provide an unstructured data pipe. Since VPs and VCs may be switched in the ATM network, the VPI/VCI at different ends of a connection may not be the same. At each link between ATM nodes, the VPCI is explicitly translated and re-assigned.

A limited number of VPCIs on a port may restrict the number of applications that can run simultaneously over that section of the ATM network, so maximum number of VPCIs is an important consideration in selecting a switch. The Bay Area Gigabit Network (BAGnet) employs an ATM service offered by Pacific Bell, which uses early Newbridge switches. Running out of VCIs proved to be a problem for interconnecting hosts in a fully meshed configuration, so for them, the number of available VPCIs proved to be a limitation.

VPIs and VCIs are not addresses. VPCIs are temporarily assigned for multiplexing, demultiplexing and switching a cell through each leg of its trip through the network. For global addressing, ATM uses 20-byte NSAP (Network Layer Service Access Point) schemes to uniquely identify user ports.

VCs and VPs are always, by definition, unidirectional. When allocating VPCI values to a Virtual Connection, the same values are used in both forward and backward direction. A virtual connection can be said to consist of two virtual channels (one in each direction). Whether or not both directions are used then depends on bandwidth allocation for the virtual channels.

A Call is an end-to-end association between ATM endpoints. A Virtual Channel Connection (VCC) is a connection between two neighboring entities in the ATM network (hop-to-hop). In this model, a call may have more than one connection. One way to implement a call might be to assign a VPI and to translate the VCIs at each hop while maintaining the VPI. Translation of VPI/VCI may be necessary if that combination is already used at a given port.

Performance

3.6 Switch capacity

Aggregate switch capacity is the maximum number of bits that can pass through the switch fabric per second. Aggregate port capacity is the maximum number of bits that all ports on the switch can feed into the switch, with continuous traffic running on all ports at the highest supported data rate.

Some vendors claim a higher switch capacity than all the ports in a maximum configuration can feed in. While it's possible the switch can pass that number of bits internally, it cannot possibly put out more bits than it takes in, so switch capacity values may not reflect the actual bitrate coming out of the switch, not to mention additional delays in output buffers.

There is another common miscalculation in deriving the switch capacity from the port capacity. SONET lines are bidirectional[1]; that is, there is one fiber for incoming and one for outgoing, and so in theory the "capacity" of an OC-3 port is 310Mbps (twice the 155 Mbps line rate). If a switch can drive all 16 ports in both directions, it would seem that the switch's "capacity" is really 16 x 310Mbps = 4.96 Gbps. The flaw in this reasoning is that the input traffic of one port forms the output traffic of another, so that should not be counted twice in computing the aggregate throughput.

Aggregate switch capacity and aggregate port capacity should be looked at together. If aggregate switch capacity is less than the aggregate port capacity, since then it becomes the upper bound on maximum throughput. If aggregate switch capacity is greater than aggregate port capacity, this may possibly predict speed scalability of the architecture for faster ports, or it may indicate an architectural limitation that requires increased internal speed.

An example of confusing "switch capacity" claims is the claimed capacity of two leading ATM switch vendors, Fore and Synoptics. Both support configurations of 16 OC-3 ports, for an aggregate port capacity of 2.5 Gb/s. However, Fore's fabric runs at 2.5 Gb/s and Synoptics' claims 5.0 Gb/s. With the same port configuration, how is it that the Synoptics switch is apparently twice as fast? In the Synoptics space-division switch, there is a possibility of contention and hence cell-dropping at any switch element port. To compensate for this, the internal clock rate was doubled to switch 5.0 Gbps, giving two clock cycles to arbitrate two cells

1. ATM and its VPI/VCIs are bidirectional, but the data rate in each direction is defined separately. In ATM, a call with 0 bandwidth can be set up (usually to reserve the VPI/VCI in the second direction), though in SONET, each link is bidirectional and of equal bandwidth.

vying for a port. It is accurate to say that the switch runs at an internal speed of 5.0 Gbps, but can only handle a maximum load of 2.5 Gbps with 16 ports running at 155 Mbps to remain virtually non-blocking. If the port speeds are doubled, all traffic can still be handled without changing the switch fabric, but in that case it will not be able to handle contentions at all.

Similarly, Lightstream's switch fabric runs at 2.0 Gb/s, but it can have 18 OC-3 ports, clearly exceeding the fabric capacity. Lightstream says the apparent discrepancy between switch capacity and aggregate port capacity values is handled by input buffering, and claims that its superior throughput results are due to clever buffering and traffic management, despite an apparently handicapped fabric.

In sum, maximum aggregate port capacity should be considered along with internal switch capacity, but is still by no means the final word on actual throughput. Some vendors offer a value for user throughput, which ultimately an ATM customer is most interested in for their own network. However, the methods and conditions under which such measurements are made can vary so much that this figure is of marginal value.

3.7 Transit delay

Transit delay (also referred to as "switch latency") is the time it takes a cell to enter, pass through, and exit the switch. Since wire speed has become so fast and reliable, the dominant delay in the network is the switch transit delay. Transit delay is by no means the final word on throughput, since it does not take congestion into account, though it establishes a lower bound for total latency over a connection (since the total delay will be at least transit delay). Measurements taken under different switch load and configurations have an impact on transit delay measurements as well.

Transit delay can be measured as a loopback (in and out of the same port), or from one input port to a different output port. It can also be measured as the delay between individual cells in a stream of cells. An HP ES4210 ATM test tool is excellent for testing transit delay, but at a cost of over $100,000.

3.8 Connection setup time

Connection setup time is the time between when a connection is requested, routed, and granted. In practice, a long connection setup time becomes a serious issue during any sort of recovery when circuits must be rebuilt. In the LAN, connection setup overhead may be more of an issue than in the WAN. In the local area, ATM's cumbersome channel setup procedure, particularly for connections frequently set up and torn down, is a weakness as compared to other packet-switched LANs. In the wide area, a fast, jitter-free link offsets the overhead in the connection setup overhead, even with the added time due to latency.

Connection setup time is affected by: load of call processor, distance from requestor to call control processor (local in a distributed system, remote in a LAN with centralized call control), geographical distance between endpoints (a WAN latency issue), network distance between endpoints (hops). A connection may be refused, or routed over a non-optimal path, if a requested QoS parameter cannot be guaranteed.

Interfaces/Compliance

3.9 ATM Interfaces

ATM interfaces can refer to host interfaces (an ATM adapter for a host computer, workstation or router), port interfaces (a switch port connected to a host interface on the other end, across which a UNI runs), or a network interface (switch ports connected to another switch port, across which an NNI runs). Switch port interfaces are also called "edge" interfaces, and network interfaces are also called "trunk" interfaces.

ATM service was originally designed to run over SONET-defined rates, though other signaling formats to carry ATM cells are commonly used as well. For example, network interfaces run over DS-3 and port interfaces run over TAXI. An ATM port's "speed" is determined by the physical interface and signalling on a port.

SONET's rates are defined as multiples of 51.84 Mb/s STS-1 (Synchronous Transfer Signal) channels for electrical signalling, or OC-1 (Optical Carrier) channels for the optical signalling. An OC-3 (or STS-3) channel is capable of carrying 3 OC-1 (or STS-3) channels at 155.520 Mb/s. A 'c' indicates that channels are concatenated: they operate as a single channel instead of n multiplexed channels. OC-N and STS-N are not compatible with OC-Nc and STS-Nc, respectively. An STS-N signal can be carried on any OC-M, as long as M is greater than or equal to N [1], and at N rates. An STS port can talk to an OC port with an electrical-to-optical conversion provided via transceivers.

SDH is the CCITT/ITU (international) version of SONET that defines a similar synchronous multiplexing structure, in multiples of 155.52 Mb/s STM-1 (Synchronous Transfer Module) channels. An STM-1 frame is structurally equivalent to an STS-3c frame, though minor differences make them currently incompatible. SDH does not make a distinction between optical and electrical signalling as SONET does, so there is no SDH OC equivalent.

Most vendors offer OC-3 and DS-3; TAXI and T1/E1/T3/E3 are also common. Carrier service vendors are more likely to have WAN offerings, and not all have OC-3.

OC-12 host and port interface availability is currently limited by the silicon. Many vendors are aiming for Q2 or Q3 1995 to release OC-12 interfaces. Certain vendors such as Newbridge and TRW claim OC-48 for internal switching and network interfaces between their own switches.

SONET & SDH Frame Formats

SONET	*SDH*	*Mb/s*
OC-1 / STS-1	*n/a*	*51.84*
OC-3 / STS-3	*STM-1*	*155.52*
OC-12 / STS-12	*STM-4*	*622.08*
OC-48 / STS-48	*STM-16*	*2488.32 (2.5 Gb/s)*
OC-192 / STS-192	*STM-64*	*9953.28 (10 Gb/s)*

Typical ATM Port/Network Interfaces

Format	*Mb/s*	*Physical*
T1	*1.5*	*copper/coax*
E1	*2*	*copper/coax*
E3	*34*	*coax*
T3	*45*	*coax*
DS-3	*45*	*coax, fiber*
4B/5B	*100*	*fiber*
TAXI	*100*	*fiber, UTP/STP*
TAXI	*140*	*fiber*
STM-1	*155*	*fiber, UTP/STP*
STS-3(c)	*155*	*UTP/STP*
OC-3(c)	*155*	*fiber*
STM-4	*622*	*fiber*
STS-12(c)	*622*	*fiber*
OC-12(c)	*622*	*fiber*

3.10 Non-ATM native interfaces

Some port modules on ATM switches may provide direct interfaces to LAN and WAN protocols such as Ethernet, token ring, frame relay and FDDI. Ports such as these allow the switch to behave as a bridge and switch LAN traffic in-band (mixed with ATM traffic). These ports do not send or receive ATM cells, nor do UNI or NNI signalling; rather the UNI is integrated into the native interface. Hence, these ports are not considered ATM ports. Lightstream is an example that provides "edge ports" for Ethernet, FDDI, token ring, and frame relay.

Most switches also have out-of-band non-ATM ports built in for administrative reasons that are not counted in the port count. Typical interfaces are serial, Ethernet or FDDI.

3.11 UNI Signalling

A UNI (User-to-Network Interface) is an ATM Forum standard that specifies the signalling between an endsystem (e.g. computer, router) and the ATM network. The UNI standard is currently at revision 3.0, with version 3.1 under consideration for ballot; and with new features (such as anycast and an ABR service class) under discussion for UNI 4.0.

Most switch and interface card vendors belong to the ATM Forum and are involved in signalling standards proposals. (Current membership is over 500, making it harder for them to reach agreements.) All vendors try to be standards compliant, while offering features based on prestandard methods.

UNI signalling is based on B-ISDN Q.2931 signalling, with extensions to support point-to-multipoint connections. UNI 3.1 is not backward compatible with UNI 3.0.[1] Signalling procedures include the use of addressing to locate ATM endpoints, allocation of resources (VPI/VCI, bandwidth) for user connections, and negotiation between ATM endpoints for selection of end-to-end protocols and their parameters (e.g. cell rate and Quality of Service).

Since ATM does out-of-band (not mixed in with user data) signalling, multiple signalling schemes can coexist by using different reserved signalling channels. Hence, vendors can conform to standards as they are adopted without backward compatibility issues, though realistically, supporting a protocol consumes memory and processing resources, making a product more expensive. ATM pricing is very competitive, so there is great incentive not to support multiple protocols, even though a vendor with a proprietary signalling protocol may say it is easy to change to a standard one.

Software/Features

3.12 Multicast

Multicast (multipoint-to-multipoint connections) is advertised as a major advantage of ATM. ATM switch architectures are all capable of sending out a cell from one port to multiple ports within a switch, and most vendors advertise "multicast" capability in reference to what is really an efficient mechanism for *cell replication*, not multicast. Cell replication is a basic function that enables higher-level multicast addressing methods.

One-way point-to-multipoint connections are easily implemented in ATM because of the efficient cell replication within the switches, and because only one host needs to know the addresses of the recipients. However, multipoint-to-multipoint connections are more complicated. Multicast's basic issues are maintaining a list of recipients (group address), needing a mechanism to translate the group address into a multicast distribution of data, sender- or receiver-initiated join to a multicast

1. Due mostly to referencing an updated ITU standard (Q.21X0) for SSCOP (Service Specific Connection Oriented Protocol), which is a general purpose data transfer layer providing, among other things, assured data transfer.

session, if a non-member of the group can broadcast to the group address, and how the group address is advertised.

One prestandard solution for ATM multicast is overlaid point-to-point connection trees with a multicast server administrating the connections. This has the problems of delay and failure potential in any solution that involves a central point, but may be the best solution given what is available with current standards.

There is a growing belief that ATM's multicast mechanism must be at least as flexible and robust as IP multicast, drafted in RFC 1112. IP Multicast uses "host groups," a set of hosts identified by a well-known single IP address. Multicast routers forward multicast packets to the remote networks with the host group destination addresses, and local network multicast reaches the local host group destination addresses.

The UNI 3.0 standard supports unicast (point-to-point) VC and point-to-multipoint VCs. UNI 4.0, currently under consideration, will offer features that will facilitate implementation of multicast, such as "anycast," which specifies a well-known VC and an "anycast home," an end system capable of administrating various services. Presumably anycast will have some standard solution for the robustness issue of the centrally administrated anycast home.

Multicast is necessary to support LAN emulation (transparently interconnecting legacy LANs); as well as applications such as video-on-demand, teleseminars, conferencing etc. ATM multicast is not standard yet, but many campus-type vendors provide a prestandard multicast offering anyway (Newbridge, Fore).

3.13 SVCs

With Switched Virtual Connections (SVC), ATM virtual channel connections (VCCs) are dynamically established and released as needed, as opposed to Permanent Virtual Connections (PVC), which are set up statically and through administrative procedures. SVCs are standardized in UNI 3.1, which does the signalling that sets up SVCs. Not all vendors offer SVCs yet, though all intend to.

SVC parameters are set up on-demand by an application, whereas PVC parameters are entered manually by the network administrator. For basic applications traditionally served by connectionless service (such as file transfer), PVCs are of little use. PVCs are also routed statically, not taking advantage of robust network topology in the event of switch or link failure along the route. With PVCs, the VPI, VCI and end ports are all pre-assigned and there is no switching. PVCs merely emulate dedicated paths, and are a relatively uninteresting happenstance of ATM, whose real power is in SVCs.

SVCs involve issues and software features such as shortest-path routing, topology discovery, alternate path routing and connection recovery. With an SVC, the network uses a routing algorithm to determine the shortest path of links and nodes over which the connection is routed. A topology discovery algorithm is necessary to find the topology and available routes. Alternate path routing routes a connection over a non-optimal, non-shortest path if some resource along the shortest path cannot be granted (not enough bandwidth, not enough VCIs). Connection recovery transparently re-routes an existing connection interrupted by a link or switch failure along its path.

3.14 NNIs

For an ATM network to serve as a backbone, the Network-to-Network Interface (NNI) is necessary to connect ATM switches to each other. As of this writing, the standard for P-NNI is under review, so vendors that offer NNIs must use proprietary signalling and routing, or use PVCs for interswitch communication in the interim. ATM switches may be used in local area networks with just UNIs, but without NNIs there is no network of ATM switches.

NNI links are different from UNIs since neither end terminates user connections, but may establish connections amongst themselves to exchange routing information. Instead of host interface cards, they have ATM (network) interface cards that must pass along the same signalling as the host interface cards, but don't need to do AAL functionality (forming ATM cells from user data). The NNI signalling will not differ greatly from the UNI signalling except in the differences mentioned above.

3.15 Traffic management

Traffic management allows a connection to specify, and forces it to conform to, certain cell traffic characteristics to guarantee a certain level of service. It provides the justification and the means for cell-dropping in response to congestion. Measurable and specifiable traffic characteristics are combined into various quality of service classes and contracts.

Congestion avoidance attempts to prevent congestion at any level, and is applied in many ways. Switch congestion control concerns itself with hot-spots and points of contention and congestion within a switch. Network congestion control concerns itself with the ability of the network to efficiently carry the offered load. **Congestion control** schemes address traffic that does no fall into a strict traffic management contract. **Flow**

control concerns itself with smooth end-to-end traffic flow over a given connection, to avoid the necessity of dropping cells. **Priority levels** allow higher-priority cells to be switched before low-priority cells without cell-dropping. Multiple priorities may exist within a class of service.Though these schemes are distinct from traffic management, all contribute to maintaining quality of service for a connection.

The following breakdown of traffic management aspects is a gross oversimplification of an extremely complex and interrelated topic. For a more in-depth treatment, see [1], Chapter 12.

Traffic contract

All traffic management schemes take advantage of ATM's model of a call by establishing a traffic contract between a connection and the network. A traffic contract exists for every ATM connection, and is an agreement between the network and the user that specifies Quality of Service (QoS) parameters (grouped into QoS classes), traffic parameters and a conformance checking rule. Each switch accepts or rejects a connection based on if the resulting traffic mix will allow it to achieve all its traffic contracts.

Traffic parameters

The traffic parameters for a connection in which terms QoS classes are defined are: average cell delay, cell delay variation ("jitter"), cell loss probability and cell error.

QoS classes

Specified QoS classes indicate values for certain traffic parameters (e.g. cell loss ratio < 1%). Every specified QoS class meets performance requirements for a *service class,* such as Constant Bit Rate (CBR) and Variable Bit Rate (VBR) and their derivatives. An Unspecified QoS class indicates no values for traffic parameters.

One unspecified service under consideration by the ATM Forum is Available Bit Rate (ABR). With ABR, traffic parameters are not considered and instead a flow control mechanism throttles traffic onto the network, allowing a cell loss ratio service guarantee. Rate-based and credit-based flow control schemes are under consideration for standardization, and a recent vote in the ATM Forum Traffic Management group was to support rate-based flow control mechanism for ABR traffic. Many vendors provide pre-standard ABR service today.

Another service supported by an unspecified QoS class is Unspecified Bit Rate (UBR), also known as best effort, and by some as worst effort. With

UBR, there is no mechanism to manage traffic and there is no service co commitment from the network. Instead, the user application is expected to adapt to the time-variable, available network resources and recover from cell loss.

Quality of Service Classes

QoS class	Traffic parameters	Service class	Application
0	unspecified	UBR, ABR	best effort
1	specified	CBR	voice, circuit emulation
2	specified	VBR	video, audio
3	specified	conn. data	data transfer
4	specified	conn.less data	data transfer

Conformance checking

Conformance checking is any algorithm that checks a cell stream against the traffic parameters set up in the traffic contract. ATM cells contain a Cell Loss Priority (CLP) bit. Traffic management schemes can use the CLP bit to tag lower-priority cells as targets for dropping, buffering or treating the aggregate of tagged (CLP=1) and untagged (CLP=0) as a separate cell stream.

The ATM Forum UNI standard describes the *leaky bucket* algorithm. A leaky bucket algorithm examines an arriving cell stream for cells not conforming to its traffic contract and discards them. Nonconforming cells are dropped or tagged for dropping by setting its CLP (Cell Loss Priority) bit to 1.

A dual leaky bucket algorithm tags nonconforming traffic and send it to a second bucket, which may or may not drop the cell. Each bucket's traffic contract determines how it identifies and treats a nonconforming cell, and may not be the same for the two buckets.

For instance, the first bucket may check the untagged cell stream (CLP=0) for Sustained Cell Rate (SCR) conformance and tag nonconforming cells, and the second checks the aggregate tagged and untagged (CLP=0+1) cell stream for Peak Cell Rate (PCR) conformance.

Traffic contract parameters

The ATM Forum has agreed on several traffic contract parameters, along with leaky bucket conformance checking, as described below. Traffic contracts are formed by choosing or combining these parameters.

PCR: Peak Cell Rate (peak)
– minimum intercell spacing

CDV: Cell Delay Variation

– maximum number of back-to-back of cells sent at line rate (tolerance set by network)

SCR: Sustained Cell Rate (average)
– maximum average rate cells can be sent at peak rate (i.e. maximum burst size at PCR divided by minimum burst interarrival time)

MBS: Maximum Burst Size (burst)
– maximum number of cells that can be sent at peak rate

Traffic policing

Also known as Usage Parameter Control (UPC), traffic policing is the action taken by the network against nonconforming traffic (the law enforcer, if you will). UPC implementation is not standardized, but rather specified in relation to the standard leaky bucket algorithm. UPC is the function in the leaky bucket algorithm that tags, monitors and discards cells.

Traffic shaping

Traffic shaping is performed by the user application or in host adapter cards to force a cell stream to conform to its network traffic contract.

Priority pre-emption is a traffic shaping function that selectively drops cells. It is different from traffic policing in that it drops cells at the point of congestion based on traffic type (e.g. CBR), whereas traffic policing drops cells based on contract conformance. Using buffers at the edge of the fabric, priority preemption can empty a saturated queue instead of dropping cells.

Selective frame loss, or *early packet discard*, is a traffic-shaping function performed in a switch on an edge port. If a host exceeds its traffic contract, an entire frame or packet is dropped rather than converting it to cells and sending it into the backbone network, where some cells would likely be dropped. In either case, the frame requires retransmission, but selective frame loss reduces cell congestion from non-dropped cells in a frame, and hence is effective as a congestion control strategy.

Congestion control

Congestion control is needed for traffic not strictly controlled by traffic contracts, as is CBR and VBR traffic.

Congestion can seriously degrade and even cripple network performance. For example, congestion at a switch port might translate into delayed NFS traffic, resulting in NFS timeouts, retransmissions, more cells sent, more congestion and delay; eventually leading to a state similar to one known in TCP as congestion collapse.

Cell loss in an ATM network is caused primarily by cell-dropping within switches when there is contention and hence congestion. A single lost cell can result in a lost burst or packet to the higher-level protocol, so a small cell loss rate may result in much higher data loss and retransmission for the application.

Congestion control schemes often employ traffic management functions, and the following "congestion control" schemes will sound familiar after reading about traffic management. Some congestion control schemes are outlined as follows.

In a *Peak Bit Rate* allocation scheme, the user (a connection request) indicates the maximum rate at which cells are sent to the network, so that the sum of peak rates of VCCs over a link does not exceed the link's maximum cell rate. In a *Minimum Throughput* scheme, a minimum throughput requirement is specified, which may be exceeded, but the connection is refused if that minimum rate is cannot be guaranteed.

Fully booked *Connection Admission Control* admits only connections which have traffic parameters that will not cause degradation of QoS on other connections (something of an aggregate worst case scheme forbidding oversubscription of resources).

Fast Buffer Reservation manages bursty traffic with high peak rates by requesting buffer space in a switch before a burst of cells is admitted, ensuring that no burst is lost due to lack of buffer space for single cells belonging to that burst. Fast Buffer Reservation is an example of a scheme that sits the fence between congestion control and traffic management.

Selective cell discard is much like a traffic policing function, but cells are not dropped based on traffic contract conformance. Selective cell discard can drop cells destined for a particular port, or forbid cells access to the switch fabric, depending on the congestion level in the switch and the cell's traffic and/or Cell Loss priority. An example of a selective cell discard scheme is to deny CLP=1 cells of QoS levels 2 and 4 access to the switch fabric when congestion level is low, then when congestion is medium, deny port access to all CLP=1 cells first, then to QOS level 4 cells.

<u>Congestion avoidance</u>

Congestion avoidance attempts to maximize usage of network resources, even oversubscribing them somewhat, without reaching breakdown due to congestion.

Backward/Forward Explicit Congestion Notification (BECN/FECN, a flow control method borrowed from Frame Relay) sends explicit control messages to the source/destination indicating congestion. A source may then squelch new traffic; or a destination may signal the source to perform traffic shaping or error control measures. This scheme does not drop cells.

Call Admission Control overbooking slightly oversubscribes resources by admitting more connections than the traffic parameters would indicate can be supported and still maintain QoS for all connections. This scheme depends on a predictable statistical nature of the traffic, and must be used in conjunction with a congestion recovery scheme.

Call Admission Blocking refuses a connection altogether if its admission will prevent meeting other traffic contracts, and is widely used for connection-oriented services.

Flow control

Flow control methods are also used for congestion avoidance. Well-known flow control schemes are *window-based*, *rate-based* and *credit-based* flow control. An ATM forum workgroup settled on rate-based flow control for congestion control for non-contract traffic.

Summary:

Traffic management: establishing and enforcing a traffic contract.
Traffic contract aspects:
- QoS classes (e.g. CBR, VBR)
- Traffic parameters (delay,loss, jitter, error)
- Conformance checking (leaky bucket)
- Traffic contracts:
  - PCR: Peak Cell Rate (peak)
  - CDV: Cell Delay Variation
  - SCR: Sustained Cell Rate (average)
  - MBS: Maximum Burst Size (burst)

Traffic policing:
- UPC: tagging and dropping cells in the n-leaky bucket algorithm

Traffic shaping:
- Priority preemption

Selective frame loss

Congestion control for non-contract (ABR, UBR) traffic:

Peak Bit Rate allocation
Minimum Throughput
Connection Admission Control
Fast Buffer reservation

Congestion avoidance:

Backward/Forward Explicit Congestion Notification(BECN/FECN)
Call Admission Control & Blocking
Flow control

3.16 LAN Emulation (LANE)

LAN Emulation transparently interconnects legacy LANs, performing protocol conversion, address resolution and requiring multicast/broadcast delivery. It is an important feature to potential ATM users with a large investment in a legacy LAN infrastructure for two major reasons.

1) Allows an ATM network to be used as a LAN backbone for hubs bridges, switching hubs, Ethernet switches, and the bridging feature in routers.

2) Allows endstations connected to legacy LANs to communicate through a LAN-to-ATM hub/bridge/switch without requiring traffic to pass through a more complex device such as a router.

LAN Emulation does not replace routers or routing, but rather provides a method of transparently interconnecting legacy LANs. The aspects of LANE standardization under discussion are multicast, MAC address (link layer protocols such as Ethernet, FDDI) to ATM translation and SVCs. Since almost all LAN protocols depend on broadcast or multicast packet delivery, an ATM LAN must provide the same service. The LANE should be able to set up SVCs on demand and multiplex LAN traffic on existing SVCs with an ATM endpoint. LAN addresses must be resolved to a destination address (e.g. an IP address) and an ATM address for the terminating switch.

LANE requires a LANE client and a LANE server. The client can be a host adapter card, an ATM interface on a router or any ATM access device with LANE intelligence built into it. If a LAN PDU's (say, a packet's) destination is local, that is, within the LANE's immediate domain, the LANE forwards the packet locally. If the destination is a subnetwork across the ATM network, the client LANE interface converses with the LANE server to request an SVC, perform address translation, then sends the packet over ATM to the destination address.

The LANE server may occur in the every switch (distributed) or in a central controller (centralized). Also, some LANE clients may not be intelligent and have to go to the ATM network to resolve local addresses.

Examples of LAN Emulation clients are Synoptics' EtherCell and Fore Systems' LAX-20 LAN Access device. The Ethercell multiplexes and converts Ethernet frames into ATM cells from 12 10BASET ports to one ATM port. If an Ethernet frames arriving on an Ethercell ports is destined for another Ethercell port, it must go out to the ATM network to the LANE server for the ARP. The LAX-20 converts Ethernet, FDDI and Token Ring PDUs into ATM cells and does the LANE as locally as possible; i.e. it goes out to the LANE server on the switch only if it determines that the packet is destined for a remote subnetwork.

One thing to note is that for interoperability, these devices should come from the same vendor. The features to support LANE aren't standard yet, even though the protocols they're switching are.

3.17 Network Management

SNMP (Simple Network Management Protocol) has become the industry standard for a network management protocol embedded in a network device. In theory, a device with SNMP will be manageable by any network management application that also talks SNMP. In that case, if a site already has such an application, the availability of one from the vendor isn't critical, as long as the device itself supports SNMP.

3.18 Extras

Virtual LANs allow a network administrator to logically subdivide ATM endstations into logical LANs, useful for security and administrative reasons.

IP-over-ATM according to RFC 1577 specifies methods to transmit IP datagrams and resolve IP addresses to ATM addresses by use of an ATM Address Resolution Protocol (ATM ARP) service over AAL 5. An ATM network is treated as a logical IP subnet (LIS), with ATM as a direct replacement for the traditional wires, local LAN segments and routers connecting IP endstations. Some switches contain an ARP server for IP support.

Load balancing over multiple links spreads traffic across multiple links between two neighbor switches. Note this is different from link striping, in which traffic from one connection may be distributed across multiple links cell by cell. Load balancing's resolution goes to the connection level, by routing a new connection across the least loaded link.

An **Application Programming Interface** (API) allows ATM users direct access to ATM services. For example, an API allows a user to set up and tear down a call from a user application, or provides direct access to the AAL, or allows direct access to cell stream flow with no processing or error checking. An API allows a user to optimize the ATM network for a particular application.

3.19 Cost

ATM switch prices depend entirely on the port configuration. Vendors are generally reluctant to distribute price sheets, so the best metric for comparison is a typically populated configuration, such as 16 OC-3 ports for a campus switch. For the purposes of evaluation, at a later date the vendors will be asked to provide a range of cost so they are not forced to reveal competitive information, but still give the prospective purchaser a rough idea of the vendor's target market.

Roughly, ATM switches for LAN-type applications run under $100,000. Enterprise switches run up to and around $300,000. Carrier switches can go up as high as $1,000,000.

4.0 Experience

4.1 ATM Advantages/Disadvantages

<u>Advantages</u>

Fixed-sized cells allow efficient switching in hardware
Multiplexes multirate multimedia traffic
Dynamic bandwidth management/allocation
No shared media: statistical multiplexing
QoS class support, guaranteed service
Speed scalable
Distance scalable
Automatic configuration:
- topology discovery
- failure recovery
- dynamic re-routing
- load balancing

<u>Disadvantages</u>

High overhead for data transfer
Connection-oriented model complicates connectionless service
Existing network IP infrastructure may not interconnect well
Immature technology (e.g. congestion problems)
Multiple types of traffic may not aggregate well

4.2 Is ATM really useful?

This question is the subject of ongoing discussion in the high-speed networking community, some of whom say ATM's current 155 Mb/s no longer qualifies as "high-speed." Regardless, ATM is with us, and the following is a cursory examination of ATM's usefulness in various situations.

<u>In the Internet:</u>

In a talk given by Van Jacobson at the August 1994 High-Speed Networking Symposium held by Usenix, he warned of ATM's limitations as an Internet infrastructure. The typical ATM user with customer premises equipment (CPE) would use it on a much smaller scale, but those limitations are worth bearing in mind. Inherent problems in traffic-scaling the call model, reliability of PVCs in the event of node or hop failure and boundaries between the network and hosts were Jacobson's main points against using ATM, particularly to replace IP in the existing Internet.

ATM opponents assert that its connection-orientedness is one of its weaknesses, not its strengths. Connection-oriented services are commonly available over connectionless networks, usually implemented in protocols such as TCP. ATM is the reverse situation: a connection-oriented network that must still provide connectionless services. For example, this fundamental difference is the main cause of the difficulty in mapping IP multicast to ATM multicast. Jacobson points out that a "connection" is too high-level an abstraction: it constrains what services can be supported by an infrastructure. Connections work best when the call lifetime is long compared to the call setup time, which may not be the case for some data traffic, such as bursty traffic generated by keyboards. Of course, IP was specifically designed to run datagrams over a connectionless link layer protocol, so it's not surprising that ATM is not the best way to carry IP.

For trunking and interconnects, ATM is cheap and flexible. Most everyone seems to agree that the bandwidth independence of ATM makes it useful for host and network interfaces. For example, there is a commercial ATM network that has been in operation for over a year in Canada, which does nothing but switch Ethernets over PVCs. It is simple, cheap and effective for this application. Connectionless services are best for handling data (the principle at the heart of the packet switch revolution) and must be implemented in ATM networks with methods for packet encapsulation and address resolution, which support features such as LAN Emulation, IP over ATM, and multicasting.

Though ATM is sometimes criticized for its limited VPI/VCI addressing. VPCIs are not addresses. Rather, they are merely local, temporary assignments over a given hop. ATM incorporates Network Service Access Point (NSAP, ISO 8348) schemes, a well-defined system of hierarchical global addressing. NSAP schemes use 20 byte addressing, even more extensive than IP's 32 bits. When a LAN is bridged over to ATM, the 32-bit IP addresses are to be mapped to the 20-byte NSAP address.

The concern that ATM may replace IP may be on the outer fringe of paranoia, since IP on the Internet is well-entrenched and its replacement would have years of experience and fine-tuning to live up to. On the other hand, now that the so-called Information Superhighway is a household word, decisions may no longer be made based on the merits of the technology alone.

<u>In the local area:</u>

ATM will provide a good backbone for legacy LANS such as FDDI and Ethernet. It is ideal for a “collapsed backbone” which interconnects smaller LANs and ATM-access routers at the building or campus level. When a shared-media LAN segment runs out of bandwidth as users are added, a legacy LAN requires redesign or re-engineering. With an ATM backbone, it is trivial to add more bandwidth by adding in more links between nodes. This does not address bottlenecks at the LAN access points, but the backbone is configurable and easily upgraded as the demand for speed and connectivity increases. However, the robust connection setup scheme may be costly in the local area, where the QoS flexibility may not be as critical as in the wide-area.

<u>In the wide area:</u>

One of the purported advantages of ATM is scalability from LAN to WAN with no changes to the technology. However, one problem is that the model of a call or connection doesn’t scale well to distances, particularly for variable bit rate and connectionless data services, due to latency and reliability issues.

Though ATM’s cell-switching concept may scale well to wide-area, its traffic management functions may not. For example, bandwidth reservation for CBR traffic may result in underutilization of resources given longer delays due to propagation (thus increasing the importance of an as-yet unstandardized Available Bit Rate service). In the wide area, the network limitation shifts from bandwidth to delay. An application that performs well in the local area may decline when the wide-area propagation delay introduces unacceptable latency.

LANs try to support the traffic the machine wants, whereas traditionally, the WAN has supported the traffic the customer is willing to pay for, and the sources are forced to match that. This basic difference in philosophy may be more difficult to scale than the technology itself.

<u>To the desktop:</u>

ATM to the desktop is not expected for at least another year, partly because workstations can’t keep up with ATM yet. Also, legacy LANs are based on connectionless broadcast mechanisms that must be emulated in ATM for internetworking, and are not standardized yet. When ATM does arrive, some are concerned about the single-point-of-failure nature of a switch compared to the robustness offered by dual-homed FDDI. FDDI is still the fastest and most reliable medium to make it to the desktop.

<u>Multimedia applications:</u>

Multiple QoS class support, bandwidth reservation, dynamic call setup and eventual multicast service make ATM well-suited to multimedia applications. However, a concern raised by Craig Partridge in a keynote address at the 1994 High-Speed networking Symposium was that voice, video and data traffic won't aggregate well.

Partridge's assertion is that ATM traffic doesn't follow a Poisson traffic model, as telephone traffic does. The Poisson model says that over time, bursts will balance out into a uniform pattern. But data and bursty traffic (such as digital video) remains bursty over time, following a self-similar (fractal) pattern, and does not converge to an easy analytical Poisson model.

Most traffic analysis is done based on a Poisson model and hence may not be a valid indication of how ATM networks and switches will actually perform. This points to the importance of buffers and traffic management methods for optimizing switch and network throughput.

High-speed applications:

For data-intensive applications, will ATM come through?

ATM overhead is 12.8%, for a maximum effective rate of 135.63 Mb/s over STS-3c and 542.53 Mb/s over STS-12c[1]. Higher-level protocol overhead, connection setup time, switch latency and traffic management cell-dropping strategies all work to further degrade throughput.

However, for perspective, FDDI uses 4B/5B encoding, which accounts for 20% overhead, plus MAC (Medium Access Control) overhead. Ethernet's Manchester encoding costs 50%, plus MAC overhead. These make ATM's 12.8% look like a bargain.

ATM's greatest advantage for high-speed applications is its mechanism to allow users to request a quality of service, guaranteeing a certain level of throughput. In a homogeneous network dedicated to high-speed applications, in which the application does not have to compete with other users for resources, an application might request all 155 Mb/s and get up to 135 Mb/s throughput. Eventually, with OC-12, the best case will be 542 Mb/s. For some applications demanding high throughput, such as videoconferencing, ATM may be the high-speed answer. For applications demanding 400 M*Bytes*/s, ATM isn't there, but then, noth-

1. The rate of the STS-3c or STM-1 is 155.52 Mb/s. Of the 2430 bytes in a SONET frame, 27 are section overhead, 54 are line overhead, 9 are path overhead, and 2340 are payload (3.7% overhead). Thus the payload rate for STS-3c is (2340/2430) * 155.52 Mb/s = 149.76 Mb/s. The entire ATM cell, header (5 bytes) and payload (48 bytes), is mapped into the payload of a SONET frame, so the actual ATM payload rate is (48/53) * 149.76 Mb/s = 135.63 Mb/s (12.79% overhead). Over STS-12c (622.08 Mb/s), the SONET payload rate is 599.04 Mb/s and the ATM payload rate is 542.53 Mb/s.

ing is. A flexible API may be important for multimedia and high-speed application writers to best take advantage of, or work around, an ATM network's offerings.

ATM was designed with the idea to invest some overhead and inefficiency in small cell sizes for gains in running over very fast media and long distances with minimal buffer latencies. Though ATM is commonly criticized for its overhead, the lack of store-and-forwarding delays may prove to be the critical factor in throughput for high-speed applications, particularly over distances.

<u>Seamless ATM interworking:</u>

One of the applications envisioned for ATM is that it can be used from the desktop to the LAN, MAN and WAN, using the same technology end-to-end. However, there is much more to scaling ATM than just the scalable cell-switching. Applications, protocols, interfaces, methods must all scale for distance, speed, network density, and number of users. A good LAN solution for the problem of, say, connectionless broadcast service over all cells that contain a given VCI may not fit into a WAN. Problems with aggregating voice, video and data traffic may also preclude ATM as a single technology. Connection setup time over the wide-area may be seen as an investment in a jitter-free link, but wasteful over reliable, bandwidth-plentiful local media. Address resolution schemes must scale to a virtually unlimited number of users. There must be enough VPCIs per port for an increased number of connections. Clearly ATM has a long way to go before it can be a single, integrated technology to handle all networking needs.

4.3 NAS AEROnet experience

Here at NAS, an ATM prototype network has already been deployed for experimentation and testing in AEROnet, NAS's wide-area network. The primary motivation for this was cost, since purchasing wide-area ATM service from a provider offers more guaranteed bandwidth for a lower cost than for T1/T3 and DS-3, and the cost is likely to continue to decrease.

The prototype network was built on the existing network infrastructure, entailing an ATM service purchased from AT&T, three Stratacom enterprise-type switches for edge nodes (demarcation switch), and several Fore switches to connect to hosts and other local-area networks. Initial performance results showed that ATM over DS-3 was able to reach its potential maximum throughput of 34 Mb/s. Though this is less than the 45 Mb/s maximum of straight DS-3, it is cheaper and hence cost-effective, and the ATM is in place to upgrade the DS-# in the future.

The salient administrative aspect of the ATM network is one most commonly cited: PVC setup and maintenance is tedious and cumbersome, and SVCs are necessary for a fully operational network. The Stratacom switches do not at this time offer SVCs. ATM adapter cards were found to be lacking in performance. As the ATM network moves into place, applications will begin demand features such as multicast.

Testing is ongoing on this project. For more information on the network and test plan, please refer to [18] and [19].

4.4 Deployment of campus-wide ATM network at Ohio State

At a recent meeting of a SF Bay Area system administrator's association (BayLisa), Steve Romig of Ohio State University spoke about his experience deploying an ATM network to replace a large Ethernet LAN. The network connects Ohio State's computer science department's computers together and has no high-speed applications or users. So far this is one of the largest local ATM networks deployed, and the experience is worth noting.

Original network, spread across 3 buildings:

50+ fileservers
500+ workstations, mostly diskless clients; various printers
80+ trees of Ethernets
3 routers, 15 Ethernets connecting fileservers to routers
Spread across three buildings

ATM network:

23 Synoptics LattisCell 12-port fiber & UTP switches
3 EtherCell Ethernet-to-ATM multiplexers

The ATM network was intended to be slowly transitioned in by replacing Ethernet concentrators with Ethercells, which transparently added the ATM network as a backbone. The ATM backbone was heavily interconnected for robustness.

The advantages of this setup were:

- Full 10 Mb/s bandwidth from each EtherCell port
- Easy, transparent scaling of ATM backbone. If a link was too busy, simply add another link between two switches, and the bandwidth is automatically doubled.
- LAN Emulation, allowed multiple Ethernets across ATM backbone

- Virtual LANs, could replace cable plant but maintain network arrangement
- Can put switches anywhere, don't worry about router topology or protocol
- Can boot client off any server
- Automatic reconfiguration, easily rearranged (no router re-configuration)

The disadvantages and problems encountered:

- ATM networks are difficult to debug, no tools like a packet sniffer
- Diskless booting of Ethercell took some time to work:
SVCs recreated transparently if it goes away, but got into a catch-22 where can't restart daemon to create SVCs they're on a Sun, need SVC to communicate across network to get daemon code.
- Clients on EtherCells couldn't boot from ATM-attached servers:
Congestion at the Ethercells (sending from fast to slow) resulted in NFS timeouts while loading kernel. Solution was to throttle back the servers to sending 10 Mb/s.

Ohio State's ATM consideration checklist:

<u>Reliability</u>

Recovery time

Host reboot -- are SVCs recreated automatically? (yes)

Lose/regain host/switch link/switch -- alternate paths?

<u>Interoperability</u>

Does all equipment work well together (recommends using the same vendor)

Host adapter card availability

<u>Congestion control & performance</u>

May have to throttle back hosts

Junctions of slow & fast were a real problem

Not using network now, waiting for equipment with congestion control

VC setup time an issue

<u>Installation & planning</u>

Topology considerations: path redundancy, multiple links

Cell sniffer for testing would be good

Allocate for expected maximum usage, or oversubscribe to ensure big enough pipe? Not certain how to use ATM's flexibility.

5.0 Conclusions

For applications that demand the highest throughput, such as NAS's Distributed Virtual Windtunnel, ATM will not be sufficient for years. However, for many other uses, it should be a consideration. ATM will soon be widely available, and considering the enormous investment companies are making in ATM, it will likely be with us for a while.

Some of the unknowns are how it will perform with real user traffic, particularly with data. One of ATM's strengths is that it can multiplex voice, video and data; but if all we need is data, proven 100 Mb/s technologies already exist (such as FDDI). Its flexibility, easy configuration, capability to guarantee a quality of service and the exploding availability of products are all important advantages.

For choosing a switch vendor, the following key aspects should be considered.

- Type and target market
- Switch transit delay
- OC-12 availability
- Robustness
    - Switch fabric redundancy
    - Distributed call control
- Traffic management
    - QoS classes
    - Buffer sizes & management scheme
    - Congestion Control
- NNI, SVC a bare minimum
- IP-over-ATM support
- LAN Emulation if to be interworked with LANs
- Adapter card availability and proven interoperability

6.0 Vendors

In the campus arena, Fore is the most visible contender and is regarded as the industry leader. Fore's ASX-200 switch architecture is a 2.5 Gb/s TDM-bus, uses distributed call control and standard traffic management features including shared buffer pools. They are working on offering more than 16 ports, but the TDM bus will eventually limit how many ports can be added. Fore excels in LAN features such as LAN emulation, with an ARP server built right into the switch, and an intelligent LAN access box that doesn't route local traffic through the switch. Fore also does well in performance tests, showing a lower transit delay than multi-stage-based competitors.

Synoptics LattisCell's switch architecture is a 5.0 Gb/s multistage matrix with centralized call control and a maximum of 16 OC-3ports, for a maximum possible throughput of 2.5 Gb/s. Like Fore, Synoptics concentrates on the LAN market and offers the whole array of LAN features. The literature and a Synoptics representative indicate no traffic management or congestion control as of yet; the representative citing a commitment to standards and there not being any yet. This switch is based on a well-published architecture designed by Jon Turner at Washington university, and is the only campus switch surveyed with a multistage matrix architecture. While this architecture is touted as being easily scalable for adding ports, so far it has remained at 16 (no doubt due to aforementioned problems in scaling space-division switching). Synoptics is also working on ATM modules to fit into their intelligent hubs, and may use this chassis to achieve more than 16 ports.

Newbridge is hot after the ATM market with just about every ATM offering possible, from a 9.6 Gb/s single-stage matrix carrier switch to a 1.6 Gb/s "workgroup" ATM LAN switch. Newbridge also offers what they call an Interstage card that allows transparent interconnection of one of their carrier switches as the backbone between multiple workgroup switches. Newbridge's approach to congestion control is the standard dual leaky bucket traffic management in the workgroup switch, but with traffic shaping instead of traffic policing in the carrier switch.Cisco's Hyperswitch was jointly developed with NEC, and allows up to 16 OC-3 ports. It uses a 2.4 Gb/s single-stage matrix with input and output buffers, distributed call control and two traffic priority levels, discarding CLP-marked nonconformant cells. Naturally, Cisco also offers ATM interfaces for their routers.

Wellfleet, Cisco's arch-enemy in the router world, offers only ATM product peripherals, including an ATM interface for their routers that may be useful to sites with a large installed base of Wellfleet routers. Wellfleet's ATM switch offering is through their recent merger with Synoptics.

Lightstream is a spinoff company from BBN, purchased in 1994 by Cisco Systems. The Lightstream 2020 offers a 2.0 Gb/s optionally redundant single-stage matrix architecture with a current maximum of 18 OC-3 ports. It has native interfaces for FDDI, frame relay, ethernet and others, and with its LANE and VLAN offerings, straddles the campus and enterprise market. Lightstream says it has paid careful attention to buffering, traffic management and congestion control schemes to handle a greater port aggregate port capacity than switch fabric speed and maximize throughput. They are one of the few to publish throughput values (3.2 Mcps (cells per second)).

NET was one of the first companies to release an ATM switch, and at the time they employed Peter Newman, a forerunner and frequent publisher in the area of fast packet switching technology. NET's switch falls squarely into the enterprise category, offering up to 90 ports over a backplane architecture with distributed call control. They do not yet offer OC-3, but offer the standard WAN interfaces (T1/T3/DS-3 etc.)

Telematics is aiming its 2.5 Gb/s virtually non-blocking bus-based switch at the edge node (enterprise) market, and does not expect to release a USA product until Q3 95. Its foreign customers are the German PTT and the Danish PTT.

Hughes' ATM Enterprise Network switch employs a broadcast bus fabric with the interesting feature of configurable ports. The switch can be configured between 16 OC-3 nonblocking ports and 56 blocking ports, allowing a user to assign nonblocking status to highly utilized ports that can't tolerate cell loss, and blocking status to remaining ports to allow higher density utilization. While the vendor claims a 3.1 Gb/s switching capacity, this conflicts with other information given. Access cards are connected over a 1.6 Gb/s ATM backplane to a 2.5 Gb/s switch, but the aggregate throughput will never exceed 2.5 Gb/s. It is true that 4.1 gigabits can exist in the system at one time, but this does not say anything about actual throughput. Though it does not support SVCs or NNIs, real-time adaptive routing and VC reconnect and reroute features are claimed, presumably for PVCs or future SVC support.

GDC, TRW, GTE, AT&T, NEC, Fujitsu, Stratacom and Northern Telecom are all carrier switch manufacturers, and still do not represent a comprehensive list of all carrier switches on the market. Of the carriers, the most visible and established ones are Stratacom, GDC, AT&T and GTE.

Stratacom claims to be the first one in the cell relay world with their 24-byte fast packet switch. Today they market the BPX, a broadband multi-shelf switch into which narrowband AXIS interface shelves can be inserted to convert non-ATM traffic into ATM. The BPX uses a 9.6 Gb/s single-stage fabric with 12 general-purpose slots handling 800Mb/s each, into which ATM, narrowband or broadband (to/from other BPXs) can be inserted. Though the switch is advertised as OC-12-ready, the only NNI (trunk) interfaces offered are T3/E3, so its SONET interfaces are only UNIs right now. Call control is distributed, with reroute priorities and priority bumping (but no alternate path recovery or connection recovery). Stratacom has obviously done its homework with traffic management, offering what they call closed-loop congestion avoidance and EFCN flow control in addition to the usual dual leaky bucket and connection admission control. The closed-loop scheme is a non-cell-dropping scheme that uses rate control to adjust cell admission to the network according to feedback on bandwidth utilization. Nonadmitted cells are buffered and admitted based on bandwidth availability. Input and non-shared output buffering per VC are employed, with the depth of the buffer depending on the class of service.

Stratacom has paired with Synoptics to interwork their AXIS Interface Shelf as an edge node between a Stratacom BPX carrier backbone and a Synoptics workgroup backbone. With the Intelligent Network Server, a central call control system (no doubt from Synoptics), SVCs are available. Stratacom already has several switches deployed at various NASA sites.

General DataComm (GDC)'s non-blocking 6.4 Gb/s crossbar single-stage matrix switch is arranged as 8 or 16 400 Mb/s slots, keeping 2 155 Mb/s ports per slot, for a maximum of 16 x 400Mb/s = 64.Gb/s. Small input buffers occur in the fabric, output buffers are arranged in high and low priority queues, with up to 4 leaky buckets for traffic policing. GDC's recent customer coup was MCI.

GTE's broadcast matrix switch offers fabric speed as needed, with 1.2 Gb/s per 8 ports, for a maximum configuration of 9.6 Gb/s with 64 ports. However, their literature indicates that currently, only an 8x8 switch is currently available, with plans to expand to a 16x16, which is what is listed in the vendor chart. A future lower cost switch will be expandable on a shelf basis, and will support up to 128 OC-3 ports per shelf on up to 14 shelves, and will support OC-12. The SPANnet switch uses deep buffers on the edge of the fabric for priority pre-emption traffic management instead of the usual traffic policing. It is the only switch to offer a HiPPI interface, with IP or IPI3 encapsulation. GTE says switch control is distributed *and* central. All switches have a SPARC chip and can be controlled remotely or locally, but a central control module is a single (though redundant) point of failure.

Northern Telecom markets GTE's SPANnet switch in a central office configuration, calling it the Magellan Gateway. Its 8x8 single-stage "with enhancements planned" fabric also provides 1.2 Gb/s per 8 ports, as does GTE's switch (which, according to a GTE representative, is a broadcast matrix). A mere 8 ports are currently sold, but it uses the same modular architecture that GTE uses, scalable as 8-port fabrics are added. Northern Telecom's literature was the only one which was specific with its switch's blocking probability: "Achieves non-blocking mode with a CLP of 1 in 10-9 at 0.8 random distributed loading profile per port." Four levels of priority, "dual buffers" (apparently dual leaky buckets), a large buffer for VBR traffic and a short buffer to minimize delay for CBR traffic make up its Multiple Priority System (MPS) traffic management. It uses prestandard NNIs, no SVCs, and only static routing (though dynamic is planned), central call control and is very sparse on interfaces.

The vendors who are close enough to OC-12 to be advertising or beta-testing are TRW, Fujitsu, DEC and AT&T. Most others project OC-12 with a rough date in mid-1995, and for NNIs only. OC-12 host interfaces are a long way off.

TRW provided some of the most detailed feature information of all the vendors. TRW's BAS 2010 series switches are squarely carriers, with a 12.8 Gb/s redundant crosspoint matrix architecture. TRW will offer configurations of 4, 8, 12 and 16 ATM modules, each of which can switch 800 Mb/s at full duplex. An ATM module can contain 4 OC-3 ports, or 1 OC-12 port. The future maximum configuration with 16 ATM modules is then 64 OC-3 ports and 16 OC-12 ports. The only currently available configuration is 4 ATM modules. TRW also has plans for a high-end switch to interconnect clusters of the carrier backbone switches, designed to carry 32 OC-48 trunk ports!

DEC is making ATM interfaces ("linecards") to their GIGAswitch FDDI switching bridge available, making for a fast ATM-to-FDDI bridge. DEC is also planning a CPE workgroup or high-performance LAN switch, a 10.4 Gb/s 13x13 singlestage "nonblocking" crossbar switch, with 800 Mb/s per pair of ports. DEC will not do traffic policing, but instead does some traffic shaping of CBR traffic, and plans to provide a WAN traffic shaper/filter in the future. Congestion control is in the form of credit-based flow control for ABR traffic, and eventually EFCN to support rate-based flow control. By December 1994 DEC will be beta-testing an OC-12 at Argon (national lab in Chicago), and will have OC-12 production-available in the second half of 1995. DEC's switch is not available yet, but T3 & OC-3 linecards for the GIGAswitch are available now.

Alcatel claims to be one of the first CO switches installed, which is not necessarily a competitive advantage, as NET illustrates also. With a 10 Gb/s switching matrix, the maximum throughput it can muster is 1.2 Gb/s (listed plainly in its literature). Alcatel says its matrix can process multiple cells simultaneously, allowing alternate cell routing within the matrix. Though it has 29 ports, a maximum of 8 can be used for OC-3 rates, which seems rather inadequate for a carrier-type switch. No mention is made of any sort of traffic management, SVCs, or NNIs.

AT&T added ATM modules to their Globeview 2000 Broadband system in a confusing array of configuration options. The basic Service Node has 8x8 2.4Gb/s multistage fabric, and is configurable for up to eight of these modules for a total of 20Gb/s switching capacity and 128 ports. The addition of a Services Module to a Service Node provides distributed-control SVCs. For all the glossies, the information is thin on buffering and traffic management details (as opposed to Stratacom, whose glossies are a crash course in traffic management strategies). AT&T's literature mentions signalling to indicate how much bandwidth is available, apparently feedback-based flow control. Of course, this is AT&T, so NNIs are provided. "Access Modules" are a peripheral product that multiplex services (e.g. video, frame relay, AT&T's router) to the ATM switch, but the switches themselves do not offer the array of LAN features, not surprisingly. AT&T offers OC-12 currently, but only as a trunking option.

Other vendors not surveyed are ADC Telecommunications, Fujitsu, Siemens, Ascom Timeplex, DSC, Motorola, NCR, NEC and Cascade.

Comments on the Vendor Survey

The variability in implementation of switches is large enough to make a consistent checklist of features difficult. Some of the features described in the Features section are too difficult to measure consistently or obtain information from vendors (e.g. connection setup time and transit delay), and so are excluded. However, this exclusion should not detract from the importance of these aspects in actual usage. Incomplete or blank information is due to lack of detail in literature and/or nonspecific answers from a vendor representative, and will be filled in as it becomes available. Cost depends heavily on configuration and can vary hundreds of thousands of dollars for a single vendor's product, and finding a cost normalization for comparison is a daunting task out of the scope of this paper. Cost estimates per port are available in [4], but were not verified for this paper.

Switch Vendor Chart

	Type	Fabric architecture	Fabric speed Gb/s	Fabric redundancy	Max # SONET ATM ports	Max port input Gb/s	Call control
Bay Networks	CA	MS Mx	5.0	no	16	2.5	Cent
Fore Systems	CA	TDM Bus	2.5	no	16	2.5	Dist
Cisco	CA	n/a	2.5	no	16	2.5	Cent
Newbridge Vivid	CA	SS Fabric	1.6	no	12	1.86	Cent
Hughes	CA/EN	BC Mx	2.5	optional	16	2.5	Dist
Lightstream	CA/EN	SS Fabric	2.0	optional	18	2.8	Dist
GTE	CA/EN	BC Mx	1.2/8 port	optional	16	2.5	Cent
Telematics	EN	TDM Bus	2.5	optional	14	2.2	Dist
NET	EN	TDM Bus	1.2	no	90 ❋	4.0	Dist
Newbridge 36150	EN	SS/MS Mx ✠	2.5	optional	16	2.5	Cent
DEC	EN	SS Fabric	10.4	n/a	52	10.4	Dist
Alcatel	CO	MS Mx †	10	no	29	1.2	Dist
General Datacomm	CO/EN	BC Mx	6.4	optional	32	4.96	Both
Stratacom	CO/EN	SS Fabric	9.6	yes	36 ❋	5.4	Dist ★
Newbridge 36170	CO	SS Fabric	12.8	yes	64	10	Cent
TRW	CO	SS Fabric	12.8	yes	16	2.5	Dist
AT&T	CO	MS Mx	2.4 - 2.0 ☆	yes	128	20	Dist
Northern Telecom	CO	SS Fabric	1.2	optional	8	1.2	Cent

n/a No information available.

✠ Configurable for single-stage non-blocking or multistage virtually non-blocking

† Requires verification

❋ Not SONET ports, these vendors do not yet offer OC-3.

★ Stratacom's autoroute feature is fully distributed, but SVCs require the addition of a central call control workstation.

☆ 2.4 Gb/s per 8x8 module, total 8 modules up to 19.84 Gb/s

Switch Vendor Chart (con'd)

	Output buffer cells/port	Input buffer cells/port	Traffic management	Congestion control/avoidance	# QOS classes	# Priority levels	VPCIs/ port
Bay Networks	n/a	no	none	none			1000 ✶
Fore Systems	2.4K	no	dual leaky bucket	none			
Cisco	✢	no	dual leaky bucket	none			
Newbridge Vivid	4K	no	none (shaping)	call admission control		3	2000 ✙
Hughes	4K-8K	no	policing ✢	congestion mgmt ✢		14	4096
Lightstream	4K-8K	6K	dual leaky bucket	selective frame loss	5	multiple ✢	
GTE	2K, 4K, 8K	70 ◆	priority preemption	call admission control		8	
Telematics	n/a	n/a	n/a	n/a			
NET	✢	no	no	BECN flow control		2	
Newbridge 36150	n/a	n/a	triple leaky bucket	call admission control			
DEC	128K	2500	none (shaping)	credit-based & FECN flow control			
Alcatel	n/a	n/a	none	none			4000
General Datacomm	✢	small ✢	4 leaky buckets	none		2	
Stratacom	24K	64K ❖	dual leaky bucket GCRA ☆	call admission control rate-based flow control FECN flow control ❇ selective cell discard	32		1000
Newbridge 36170	n/a	n/a	triple leaky bucket	call admission control			
TRW	✢	✢	dual leaky bucket, virtual scheduling	selective cell discard, FECN flow control		3	
AT&T	n/a	n/a	4 loss priorities	feedback flow control	4	2	4096
Northern Telecom	✢	n/a	dual leaky bucket	none		4	

✢ This feature is present but further details not provided.

✶ 4K VPCIs per slot, 16K connections total

✙ per switch, 16 VPs and 1024 VCs per port, but 2000 max per switch is the most significant value

◆ In fabric buffer, not strictly input

❖Total, allocated per VC, not per port

❇ Stratacom says EFCI, different acronym, same idea

☆ Generic Cell Rate Algorithm (Frame Relay)

Switch Vendor Chart (con'd)

	SVCs	NNIs	Multicast addressing	LANE	Virtual LAN	IP over ATM	API	Other features/comments
Bay Networks	●	●	●	●	●	❍	❍	
Fore Systems	●	●	●	●	●	●	●	
Cisco	❍	❍	❍	❍	❍	❍	❍	
Newbridge ViVid	●	●	●	●	●	●	❍	
Hughes	❍	❍	❍	❍	❍	❍	❍	
Lightstream	●	●	❍	❍	❍	❍	❍	
GTE	●	❍	❍	❍	❍	❍	❍	
Telematics	❍	❍	❍	❍	❍	❍	❍	
NET	●	❍	❍	●	●	❍	❍	
Newbridge 36150	●	●	●	●	●	●	❍	
DEC	n/a	n/a	n/a	n/a	n/a	n/a	n/a	
Alcatel	❍	❍	❍	❍	❍	❍	❍	
General Datacomm	●	●	❍	❍	❍	❍	❍	
Stratacom	❍	●	❍	❍	❍	❍	❍	
Newbridge 36170	●	●	❍	❍	❍	❍	❍	
TRW	●	●	❍	❍	❍	❍	❍	
AT&T	●	●	❍	❍	❍	❍	❍	
Northern Telecom	❍	●	❍	❍	❍	❍	❍	

Switch Vendor Chart (con'd)

	OC-12 STS-12	OC 3(c)	STS 3(c)	OC-1	STS-1	TAXI	DS3	DS1	T3	T1	*International* STM-1	E3	E1	Other
Bay Networks	❍	●	❍	❍	❍	❍	●	❍	❍	❍	●	❍	❍	4B/5B
Fore Systems	Q295	●	●	❍	❍	●	●	❍	❍	❍	●	●	❍	
Cisco	❍	❍	●	❍	●	●	●	❍	❍	❍	●	●	❍	
Newbridge ViVid	❍	●	❍	❍	❍	❍	❍	❍	❍	❍	●	❍	❍	
Hughes	❍	✶	●	❍	●	❍	●	❍	●	●	●	●	●	
Lightstream	❍	●	❍	❍	❍	❍	❍	❍	●	●	●	●	●	
GTE	Q395	●	●	❍	❍	●	●	●	❍	●	❍	❍	❍	HiPPI
Telematics	❍	●	❍	❍	❍	❍	❍	❍	●	●	●	●	●	SMDS DXI
NET	❍	❍	❍	❍	❍	❍	●	❍	●	●	❍	●	●	
Newbridge 36150	❍	●	❍	❍	❍	✜	❍	❍	●	●	●	●	●	
DEC	Q395	❍	●	❍	❍	❍	●	❍	●	❍	●	●	❍	
Alcatel	❍	✶	❍	❍	❍	❖	●	❍	❍	❍	❍	●	❍	SMDS
General Datacomm	❍	●	●	❍	❍	●	●	●	●	●	●	●	●	E2
Stratacom	❍	●	❍	❍	❍	❍	❍	❍	●	●	●	●	●	SMDS
Newbridge 36170	Q395	●	❍	❍	❍	❍	❍	❍	●	●	●	●	●	
TRW	Q295	●	❍	❍	❍	❍	●	●	❍	❍	❍	❍	❍	
AT&T	✪	●	●	❍	❍	❍	●	●	❍	❍	●	●	●	
Northern Telecom	❍	●	❍	❍	❍	❍	●	❍	❍	●	❍	❍	❍	

✶ 155 Mbps over singlemode fiber, signalling not specified

❖ 100 Mbps over multimode fiber, signalling not specified

✜ 100 Mbps and 140 Mbps over singlemode and multimode fiber, LATM (Local ATM) signalling

✪ Trunking (NNI) only

7.0 References

[1] McDysan, D.E., Spohn, D.L., *ATM: Theory and Application*, McGraw-Hill, 1995

[2] Cavanaugh, J.D., Salo, T.J., "Internetworking with ATM WANs," *Minnesota Supercomputer Center document*, anonymous ftp on ftp.magic.net::pub/magic/ip-atm.ps, Dec. 14, 1992

[3] Symborski, C., "ATM-FAQ.txt," comp.dcom.cell-relay Frequently Asked Questions file, anonymous ftp on cell-relay.indiana.edu::/pub/cell-relay/FAQ/ATM-FAQ/ATM-FAQ.txt, Apr. 24, 1994

[3] Almeda, J., Radwin, M. "ATM Testbed Project Plan for NASA Ames Research Center," *Communications and Networks Development Branch internal document*, Oct. 20, 1993

[4] Mier, E.E., "Buying ATM Gear? Caveat Emptor," *Communications Week*, pp. 66-78, May 23, 1994

[[5] Fore Systems, "ATM Switch Architectures," *FORE Systems White Paper*, Nov. 1993

[6] Cox, J.R., Gaddis, M.E., Turner, J.S., "Project Zeus," *IEEE Network*, pp. 20-30, Mar. 1993

[7] Biagioni, E., Cooper, E., Sansom, R., "Designing a Practical ATM LAN," *IEEE Network*, pp. 32-39, Mar. 1993

[8] Eckberg, A.E., "B-ISDN/ATM Traffic and Congestion Control," *IEEE Network*, pp. 28-37, Sept. 1992

[9] Turner, J.S., "Managing Bandwidth in ATM Networks with Bursty Traffic," *IEEE Network*, pp. 50-58, Sept. 1992

[10] Luckenbach, T., Ruppelt, R., Schulz, F., "Performance experiments within Local ATM Networks," *GMD-FOKUS ATMC document*: WWW@fokus.gmd.de/NHTP/conferences, Undated

[11] Burak, M., *e-mail exchange concerning ATM switch performance testing at GMD-FOKUS ATMC*.

[12] Csenger, M., "Early ATM users lose cells," *Communications Week*, pp. 1, 92, May 16, 1994

[13] Anderson, T., "A Case for Networks of Workstations (NOW)," *Presented at computing conference at Stanford*, pp. 11, August 1994

[14] Nolle, T., "Gauging ATM product compatibility," *Network World*, pp. 1, 64-65, Aug. 15, 1994

[15] Bush, S, et al., "ATM switches live up to hype," *Network World*, pp. 43-46, Aug. 22, 1994

[16] Pattavina, A., "Nonblocking Architectures for ATM Switching," *IEEE Communications Magazine*, pp. 38-48, Feb. 1993

[17] Newman, P., "ATM Technology for Corporate Networks," *IEEE Communications Magazine*, April 1992

[18] Kurak, R.S., Lisotta, A.J., McCabe, J.D., Nothaft, A.E., Russell, K.R., "Experiences with the AEROnet/PSCN ATM Prototype," *NAS Technical document (### pending) http://www.nas.nasa.gov/*, Feb. 1995

[19] Lisotta, A.J., Russell, K.R., "AEROnet ATM Service Test Plan, Version 1.0," *NAS Technical document (### pending) http://www.nas.nasa.gov/*, Feb. 21, 1995

8.0 Acknowledgements

Many thanks to Markus Burak, Rajeev Gupta and Albert Manfredi for consultation. Thank you again to Albert Manfredi for a detailed review. Thank you to Bill Nitzberg for answering an endless stream of Framemaker questions. And thank you to Alan Poston and John Lekashman for letting me do such a neat project.