RC 15572 (#69164) 3/13/90
Telecommunications 29 pages

Research Report

A Critique of ATM from a Data Communications Perspective

Israel Cidon

IBM Research Division
T. J. Watson Research Center
Yorktown Heights, NY 10598

Jeff Derby

IBM Communication Systems
Research Triangle Park, NC 27709

Inder Gopal and Bharath Kadaba

IBM Research Division
T. J. Watson Research Center
Yorktown Heights, NY 10598

A CRITIQUE OF ATM FROM A DATA COMMUNICATIONS PERSPECTIVE

Israel Cidon
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

Jeff Derby
IBM Communication Systems, Research Triangle Park,
NC27709

Inder Gopal, Bharath Kadaba
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

Fast Packet switching(FPS) is emerging as the preferred technology for future high speed, integrated networks. Asynchronous Transfer Mode (ATM) is an aspect of the Broadband ISDN Standard that is in very early stages of development. The standards activities are restricted thus far to choosing a 48 byte fixed "cell" and cell label swapping for routing. Even at this early stage, there are several concerns regarding ATM relating its suitability for data communications. These concerns are brought to focus in this paper by comparing it to an alternative approach to FPS developed at IBM called PARIS. PARIS uses variable length packets and source routing headers. By using LAN traffic data, we show that the fixed length packets in ATM can result in significantly worse transmission efficiency over variable size in many real traffic scenarios; considerably more processing power (requiring VLSI implementation) is needed to handle segmentation and reassembly overhead associated with ATM small cells, and statistical multiplexing present some unique problems. Also, we present some qualitative arguments to show that the label swapping approach for routing is more complex to implement, potentially slower in processing call setup and more difficult to support datagrams when compared to the source routing technique.

1.0 INTRODUCTION

In the current communications network technology revolution, amidst a number of ongoing debates and disagreements, there appears to be a general consensus on the following points. First, it is both feasible and desirable that all types of traffic (voice, data and video) be carried on a common backbone network; second, Fast Packet Switching (FPS) is the most suitable method to accomplish this. Here, FPS simply means that information transfer in the network is done in packets that carry additional control bits (headers) and processing in the intermediate nodes is limited to using these headers to route and schedule packets on the appropriate outgoing links. This restricted processing permits simple hardware implementation allowing one to build switches capable of handling millions of packets per second - a requirement of future networks supporting integrated services. The format of the packets and the routing and scheduling techniques are of crucial importance and forms the focus of this paper.

The CCITT has defined the asynchronous transfer mode (ATM) as the FPS mechanism to be employed in broadband ISDN (B-ISDN) [ATMA90]. ATM is based on the use of short, fixed-length packets called "cells". The IEEE 802.6 draft standard for metropolitan-area networks (MANs) employs the same cell structure as B-ISDN [IEEE88]. These standards are primarily driven by public-network providers with the goal of building integrated services public networks. The standards should permit these independently constructed networks to inter-operate easily, thereby resulting in ubiquitous worldwide service offerings. In the United States, AT&T and the RBOCs are driving the B-ISDN standard through the ANSI T1S1 committee and the MAN standard through the IEEE 802.6 committee. Initial use in US may be via SMDS (Switched Multi-megabit Data Services), a datagram service offering planned by RBOCs with interconnection of customer LANs as the primary application [SMDS89].

While the B-ISDN standard is still evolving, there is agreement on packet format and routing method [ATMB90]. As noted above, ATM packets are short, fixed-length "cells", with the overall cell length equal to 53 bytes in the current proposal. An "adaptation layer" is defined above the

---

Additional components of the network function such as end point protocols interfacing with different traffic sources as well as network control and management aspects (such as bandwidth allocation) while critical to the total network system are not pertinent to this paper.

ATM transport to map user traffic on the ATM transport, segmenting user packets into cells and reassembling cells into user packets. User packets are transported on a integral number of cells using padding bits to complete unfilled cells. This approach is very different from the traditional packet switches that use variable length and unpadded packets.

In a B-ISDN network, ATM cells are routed based on the contents of a "label" in the header of each cell. The labels are used in intermediate nodes in conjunction with routing tables to determine the outgoing link on which the cell should be transmitted. The label is valid only for the current hop and is replaced by a new label that will be interpreted at the next hop. The routing table in any intermediate node contains an entry for each label assigned on each incoming link, with the entry providing a mapping to the appropriate outgoing link and the new label to be used on that link. The assignment of labels and construction of the routing-table entries are carried out as part of a connection-setup procedure. This type of routing method is well known and used by many popular data networks (for example, SNA/APPN [BGGJP85])

While there good reasons to aim for a ubiquitous worldwide broadband network capable of providing advanced services, there are several concerns regarding the above approach. These include:

- The design is significantly sub-optimal for data. Specifically, the short cell size seems to have been motivated primarily by voice-traffic considerations, with the intent of keeping packetization delay and queueing delay small. It appears that this design choice has an adverse impact on data traffic. This issue is critical given the expectation of a dramatic increase in the volume of data traffic to be handled by broadband networks. Indeed, the first network offering to use ATM cells is likely to be one that at least initially will carry only data traffic, namely SMDS.

  Scalability to networks with gigabit/sec links. The routing mechanism as well the short cells may prove to be bottlenecks in building high end switching nodes.

These problems are more closely examined in this paper. The concerns regarding ATM approach are brought to a sharper focus by comparing it to an alternative approach to FPS developed at IBM called PARIS. We provide a brief overview of PARIS below.

The PARIS [CG88] high speed wide area networking project has been in progress since 1986 in IBM Research. A significant amount of work has been done in defining a complete network architecture for supporting high speed integrated (voice, video, and data) networks.

PARIS uses variable-length packets that can range in size from a few bytes to several thousand bytes determined by network implementation considerations such as buffer size. This variable length packet approach is referred to in this paper as Packet Transfer Mode(PTM).[2] Variable-length packets are employed in LANs and in classical packet networks. They are also used in fast-packet, wide-area network structures such as ISDN frame relay Here, it is possible to transport user packets in integral form with minimal "adaptation layer" processing.

PARIS makes use of a source routing (referred to as Automatic Network Routing or ANR) scheme which lends itself to easy hardware implementation and extremely fast call setup and take-down procedures. It also enables the implementation of additional routing functions such as copy, broadcast and transparent route switching. High packet switching performance is achieved at low implementation cost through the use of simple intermediate node algorithms which are optimized for hardware implementation. Most processing intensive tasks such as flow control, error recovery and adaptation protocols for voice and video are done on an end-to-end basis. A 4-node prototype supporting at 100 Mbits/sec links has been built to demonstrate some of the key network concepts.

Our approach to evaluating ATM is to compare it to the PARIS technology; specifically, the aspects we consider are:

Fixed slotted cell structure vs. variable unslotted packet structure

2. Label routing vs ANR

In examining the first aspect we focus on data traffic as the predominant traffic to be carried by these FPS networks. Specifically, we focus on LAN-LAN traffic since we expect this to a main source of traffic in the network. For example, the SMDS services are aimed at carrying this type of traffic initially. We consider real traffic measurement data on such LANs as well as analytical techniques to quantify the performance differences. The following issues are studied in detail:

- Data transmission efficiency

  FIFO queueing behavior

  Adaptation layer processing

  Statistical Multiplexing

---

[2] The term PTM was originated by Chuck Davin and David Tannenhouse at MIT

In examining the second main point of comparison, namely label routing vs ANR, we present qualitative arguments on the following aspects:

Nodal hardware/software complexity

- Speed and efficiency of connection setup and takedown
- Datagram support

It is important to note at this point that FPS mechanisms other than PARIS have been defined that are not based on the use of cells. One of these, usually referred to as Frame Relay, has in fact been standardized by the CCITT for use in "narrowband" ISDN. Frame Relay employs variable-length frames just as PARIS does. Unlike PARIS, Frame Relay employs label routing rather than source routing. However, in Frame Relay the use of the label is defined only at the interface between the user and the network. It is in fact possible for a network that provides a Frame Relay service to employ PARIS-style source routing internally. Given these considerations, it is possible to conclude that the results developed on the sequel can be generalized beyond the comparison of PARIS with ATM. That is, the results associated with the use of variable-length frames are applicable to Frame Relay with virtually no modification, while the results associated with the use of source routing could be applied to a network that provides a Frame Relay service while employing source routing internally.

Following is a brief summary of this paper. In the next section, by using LAN traffic data, we show that the fixed length packets in ATM can result in significantly worse transmission efficiency over variable size in many real traffic scenarios; considerably more processing power (requiring VLSI implementation) is needed to handle segmentation and reassembly overhead associated with ATM small cells. Overall, this supports our belief that, particularly for data, the offered traffic in networks will be very heterogeneous in nature with wide variations in terms of rate, packet sizes, variance, etc.. The ATM approach is to convert this heterogeneous traffic stream into a homogeneous cell stream of fixed sized small cells. The PTM approach is based on the claim that the conversion of a set of heterogeneous user traffic streams into a homogeneous cell structure introduces more overhead (in utilization and complexity) than it saves in design complexity and delay.

In the last section of the paper, we present arguments to show that the label swapping approach for routing is more complex to implement both from the processing and storage at the intermediate node when compared to the source routing technique. The connection setup/takedown may be

slower in ATM and the datagram support more cumbersome to implement. These are only qualitative in nature at this point and additional detailed comparison of operational prototypes are necessary to understand these issues further.

2.0 FIXED CELL VERSUS VARIABLE PACKET

In this section, we present and discuss various issues that arise because of the restriction to small fixed sized cells in ATM and compare it to PTM.

2.1 ATM assumptions

The basic cell structure defined by the CCITT for ATM in B-ISDN has a nominal payload of 48 bytes with a header of 5 bytes. This same basic cell structure has also been adopted for the metropolitan-area network (MAN) standard being developed by IEEE 802.6. Additionally, 4 bytes have been extracted from the nominal payload and and added to the overhead. It should be noted here that the addressing information contained in the user frame (such as IEEE MAC addresses or ISDN addresses) as well as any frame check sequence contained in the user frame are considered to be part of the payload by the adaptation layer. In terms of the cell format, the best case in terms of minimum overhead would be to assume that 48 bytes will be user payload with only 5 bytes for overhead.

Two fixed-length alternatives will be considered below, namely:

*A*53- This is the current CCITT and IEEE 802.6 proposal

*A*69- This represents an earlier CCITT and IEEE 802.6 proposal. We have included this just to contrast it with the current agreed upon standard although the differences turn out to be not very significant.

Finally, with regard to frame delimiting, it is assumed that that cell synchronization is obtained by using "synch cells" containing a unique header and that these cells are sent so rarely that their contribution to the total overhead is negligible. Since the cells are all the same length, identifying the start of each cell is then a simple matter of counting bytes.

2.2 PTM assumptions

The PARIS architecture is an example of PTM in that it employs variable-length network frames that explicitly includes adaptation-layer functions. Any user frame whose length is less than or equal to $(N_P)_{max}$ is sent in a network frame whose payload size is equal to the user-frame size. Any

user frame whose length is greater than $(N_P)_{max}$ is split into segments of length $(N_P)_{max}$. The last segment, which in general will be shorter than $(N_P)_{max}$, is sent in a network frame with payload size equal to its length. The importance of including the adaptation layer function is that implementation choices for the network can be at least partially decoupled from considerations relating to the nature of attachments to the network. For example, an $(N_P)_{max}$ of 2K bytes can be chosen based on network buffer-size considerations, without limiting the network's ability to carry bridged traffic from an FDDI ring with a maximum user-frame size of about 4500 bytes.

The variable-length alternatives that will be considered below are oriented towards the PARIS architecture in that they include the adaptation-layer mechanism outlined above as well as the notion of variable-length headers to support source routing. The specific alternatives to be considered, both with an average overhead of 12 bytes, are:

*PAR*2K- PTM with $(N_P)_{max} = 2K$

*PAR*4K- PTM with $(N_P)_{max} = 4K$

Finally, it assumed for the variable-length cases that frame delimiting is realized using HDLC flags, and that the HDLC "zero-bit-stuffing" mechanism is employed to prevent data from mimicking a flag. This will add slightly to the net overhead for any given frame depending on the bit-pattern it contains. For random data, the incremental overhead due to this bit-stuffing is negligibly small and is ignored.

2.3 Offered traffic models

For the various comparison points we want to consider, the results vary widely based on the offered traffic model. Thus, we performed studies on a wide spectrum of offered traffic models chosen from published results, extrapolations, or specific application characteristics.

In order to compare the fixed-length and variable-length alternatives with regard to data-transmission efficiency, overhead, and related issues, it is necessary to characterize the traffic offered to the network. We note that providing connectivity to LANs and MANs is expected to be a major application of fast-packet networks, with SMDS, which is targeted at precisely this application, likely to be the first available service offering in the US to employ the ATM cell concept. As a result, we have based the traffic models used in our study on reports of measured traffic patterns on operating LANs. We focus on sets of published measurements from M.I.T. [MIT86], Univer-

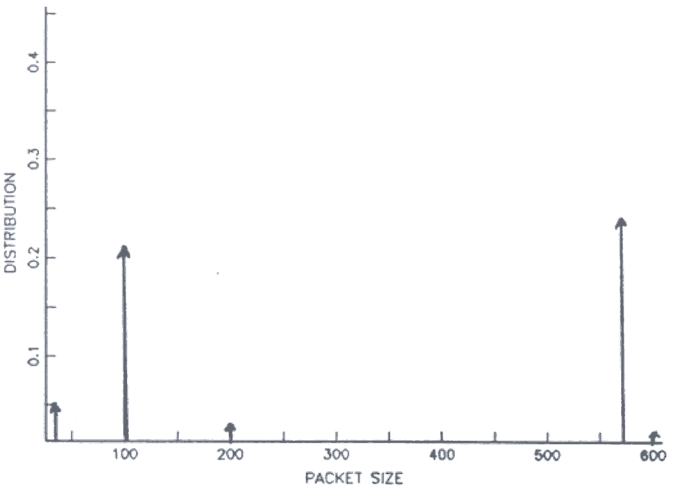
sity of Delaware [UD87] and Berkeley [BER87]. In addition, we have attempted to construct models representative of the traffic one may expect to be generated as users begin to take full advantage of wide-area, fast-packet networks. We have two sets of such models:

The packet-size distributions in the three published reports have been "stretched" by scaling their abscissas up while keeping the ordinates fixed. The result is a set of distributions in which the packets are generally longer but the relative numbers of long packets and short packets remains approximately the same.

A distribution has been created that may be representative of an application that would generally send only very large blocks of data; an image application might be one example here. In this case, the large majority of packets sent would be very long.

Our study thus considers the following traffic models:

MIT: This is the MIT distribution, shown in figure below. It displays behavior typical of several other measured packet-length distributions. This distribution is bimodal, with almost 45% of the packets sent being less than 50 bytes in length and most of the remainder being between 530 and 570 bytes in length.

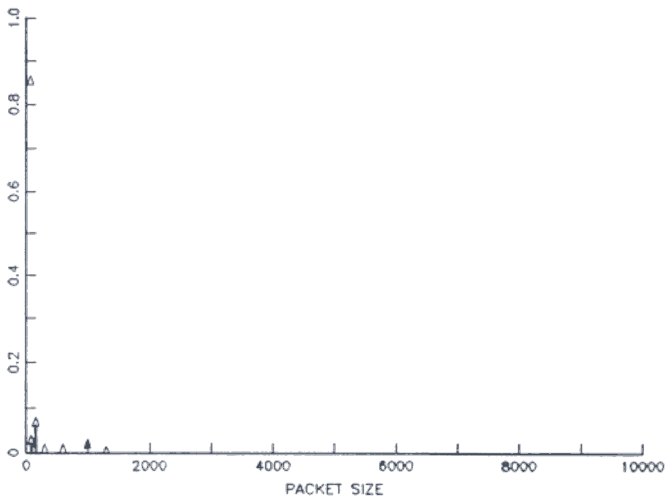


MIT distribution

We also used at the Berkeley distribution although the results are not included here for space reasons. It is bimodal, with 21.4% of the packets having a length of 46 bytes and 40.8% of

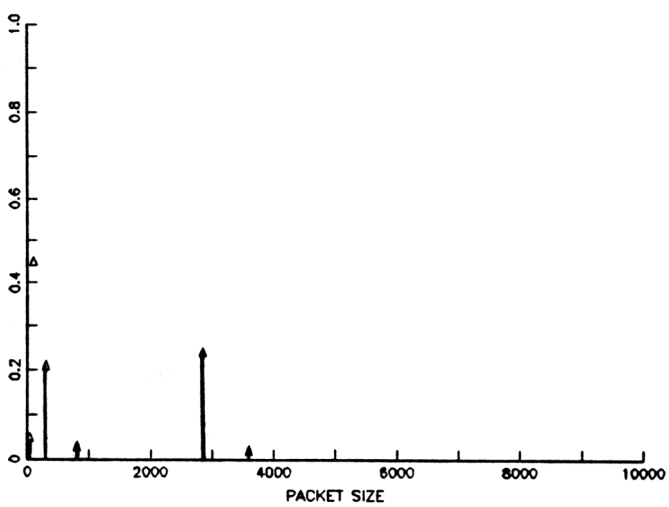
the packets having a length of 1072 bytes. In addition, almost 90% of all the bytes transmitted were contained in packets whose length was greater than 1000 bytes.

2. *DEL*: This is the University of Delaware distribution also shown below. This distribution differs somewhat from the above distributions in that in this case a large majority (87%) of the packets sent were short control packets containing 64 bytes each.

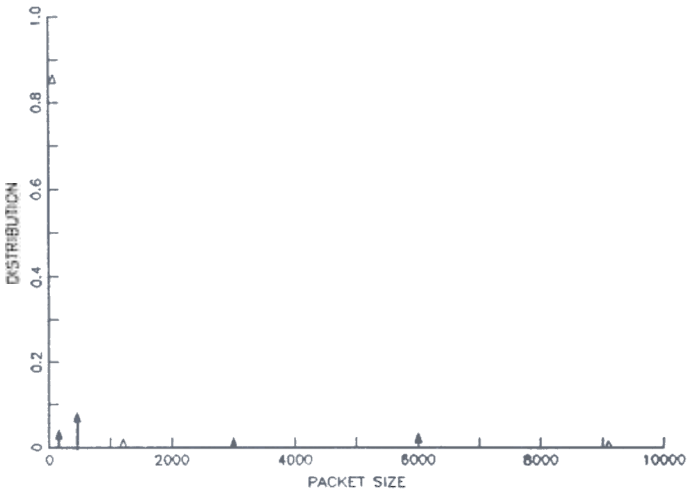


Delaware distribution

3. *MIT1*and *DEL1*: This are "stretched" version of the MIT and Delaware distributions and they are shown below.



MIT stretched distribution



Delaware stretched distribution

4. *IMAGE*: This is the distribution representing transmission of very large blocks of data. As shown in the figure, a large majority (80%) of the packets sent are about 9000 bytes in length, with the remainder assumed to be short supervisory packets. The length of the large packets was chosen to be essentially equal to the maximum packet length supported by SMDS and IEEE 802.6.

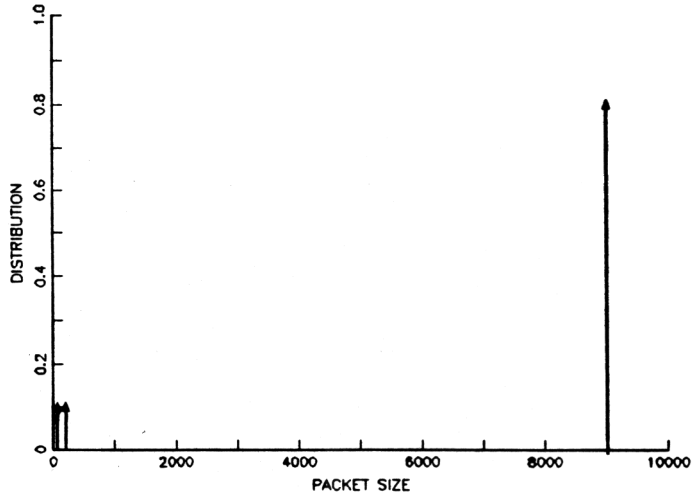


Image distribution

In using these distributions in our study, we have assumed that they specify directly the packet-size distribution in the fast-packet wide-area network. For example, in using the MIT distribution we consider a scenario in which the LAN is connected to a fast-packet WAN, with the probability that a packet from the LAN is forwarded across the WAN being independent of the packet length.

2.4 Data Transmission efficiency

The parameter of interest in this section is the data transmission efficiency. This is defined as the ratio of useful user data transmitted on a communication link to the actual bit rate of that link. The sources of inefficiency that we model are the header overhead (one header per ATM cell/PTM packet) and the cell padding caused by integrality constraints of fixed length ATM cells. We ignore additional overheads caused by packet framing, bit stuffing, cell synchronization, or SONET framing. The assumption is that these additional overheads are very specific to the underlying transmission medium and cannot easily be captured in a generally meaningful manner. Since there is only a single header to be used per user packet, PTM has intrinsically higher transmission efficiency. The use of source routing in PTM does cause a somewhat larger header size than the 9 bytes of ATM (approx. 12 bytes).

2.4.1 Analysis

Let the user data size be given by U (All sizes are assumed to be in bytes). Further let the payload in an ATM cell size is denoted by A and the header size per cell denoted by H_{ATM}. Similarly, denote the maximum PTM packet payload by P and the average PTM header by H_{PTM} The ATM transmission efficiency,E_{ATM}, is given by the expression:

$$E_{ATM} = \frac{U}{(ceiling(\frac{U}{A}) \times (A + H_{ATM}))}$$

Similarly, If all packets were of fixed size, U, the PTM transmission efficiency,E_{PTM}, is given by the expression:

$$E_{PTM} = \frac{U}{U + (ceiling(\frac{U}{P}) \times (H_{PTM}))}$$

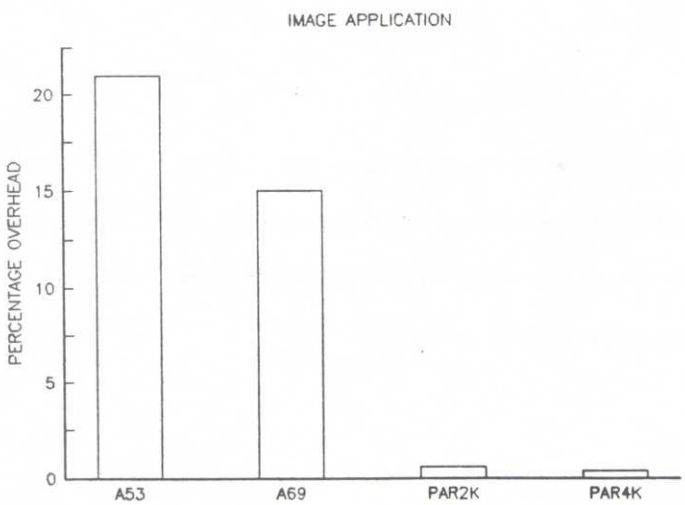
In order to calculate the average efficiency, it is necessary to use the offered traffic distribution. The average efficiency, E_{avg} is given by:

$$E_{avg} = \sum_{u} Probability\{a\ random\ byte\ comes\ from\ a\ userdata\ of\ size\ u\} \times (Efficiency\ for\ data\ of\ size\ u)$$

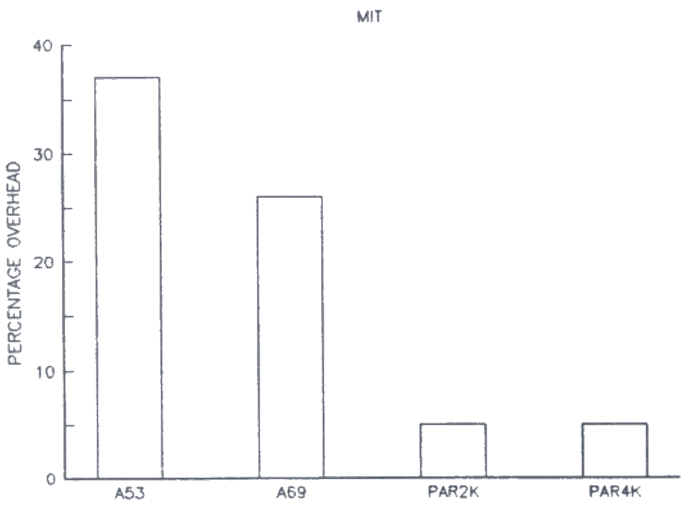
The probability within the summation comes directly from the packet size distributions. In the curves that follow, we have plotted the percentage overhead which, in terms of the efficiency, E, is simply $E^{-1} - 1$.

2.4.2 Results

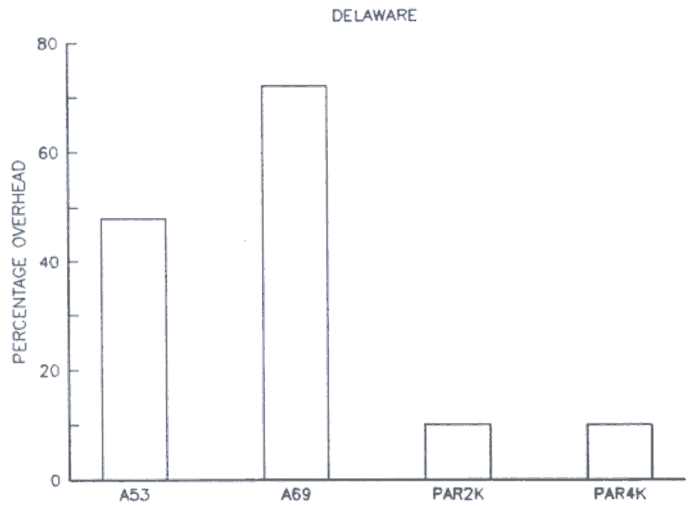
We have numerous results from each one of the distributions. To present the general trends, we have chosen four representative figures. The first case is the simulated image application. Here, most traffic is composed of 8Kbyte packets. The overheads are essentially the constant overhead (1 header/cell) of ATM compared with the vanishingly small overhead (1 header/packet) of PTM. Clearly, ATM causes inefficiency due to constant per cell overhead.



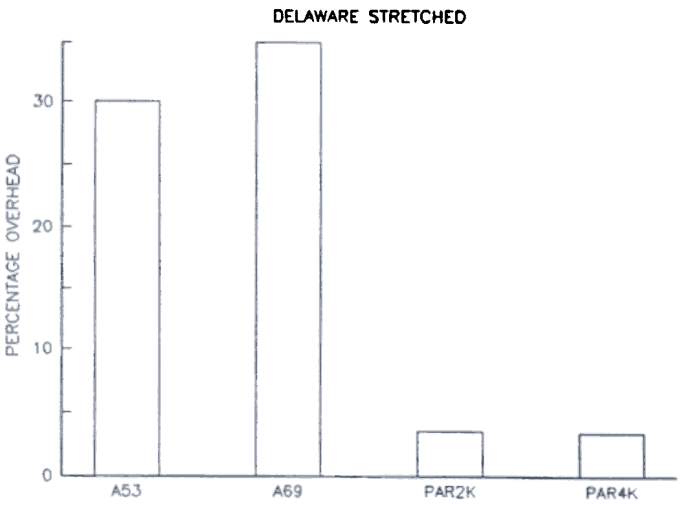
The second curve is from the MIT LAN study which was heavily bimodal in nature. Here, the second deficiency of ATM, i.e. its requirement of integral cells and the consequent forced cell padding, begins to play a role. ATM overheads are significantly more than PTM.



In the next study (U. Delaware) which is heavily concentrated around short packets, this cell padding overhead becomes more dramatic.



Finally, we have included a "stretched" version of the Delaware study. Here the packet distribution is quite widely spread out. Again, ATM performs poorly.



Conclusions

Since it is unlikely that data applications are going to be designed with the nature of the underlying transport in mind, it is necessary for the underlying transport to be general enough to carry any packet distribution without degradation of efficiency. Unfortunately, ATM does not fit the bill. In particular, it performs very poorly for certain distributions and provides efficiencies that (in extreme cases) are close to half the efficiency PTM can provide.

Buffer delays and storage

Another often repeated statement about the benefit of ATM is the low delay and small storage required in the intermediate node adaptors. The argument usually comes from (explicit or implicit) reference to a standard (say, M/M/1) queueing curve of delay vs. utilization. For a given utilization, the average delay is some fixed multiple of the service time for a single packet. If packets are smaller, the value of the delay is proportionately smaller (as is the storage requirement in terms of bytes). The purpose of this section is to point out that this argument is somewhat fallacious. In fact, in some cases, PTM delays are lower than comparable ATM delays. There are two factors (captured in this section) which make the above argument less valid. The first (less important) reason is that the arrivals of successive cells of the same user packet are highly correlated. In fact, if we observe the cell arrivals at the entry point into the network, it is typically the case that the cells of the same user packet arrive at regularly spaced intervals (the spacing being determined by the

ratio of the speed of the access line to the speed of the first link in the network). If the spacing between cell arrivals is small, the last cell in a packet is very likely to see a significant fraction of the other cells from the same packet ahead of it in the queue and consequently experience a much larger delay. The analysis that we perform in this section uses a Poisson Cluster process model proposed in [SOH89], to capture this cell bunching effect. The second (more important) reason for the breakdown of the "ATM delay" argument is that it ignores the effect of higher ATM overhead. As demonstrated in the previous section, ATM has an intrinsically higher overhead for most offered traffic distributions. To compare ATM and PTM, it is important to use situations where the carried user traffic is equivalent. As ATM has higher overhead, this results in a higher link utilization, resulting in a higher delay and storage requirement. This second effect is substantial at larger utilization, where even a small shift in utilization results in a significant increase in delay.

Analysis

The analysis makes use of the results in [SOH89]. We focus on the first node on the path and examine the delays experienced in queueing for the first link. The basic idea is to use a Poisson Cluster Process to represent the ATM cell arrivals. The relevant result is that the average number of cells in the queue, $\overline{N}$ is given by, $\frac{P \times B}{2(1-P)}$. P, is the utilization and B is a term that denotes the "batchiness" of the arrival process. The batchiness term, B, is given by the expression:

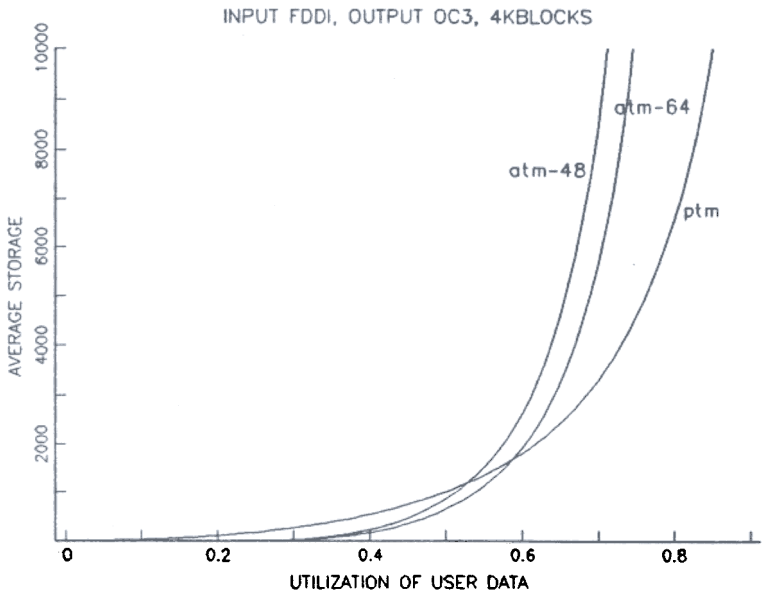
$$B = 1 + (\frac{E[g^2]}{E[g]} - 1)(1 - r(1 - P))$$

Where, r is the ratio of the incoming and outgoing rate, and g is the number of cells in a packet. In this study, we use fixed length user packets rather than the distributions used in the previous section.

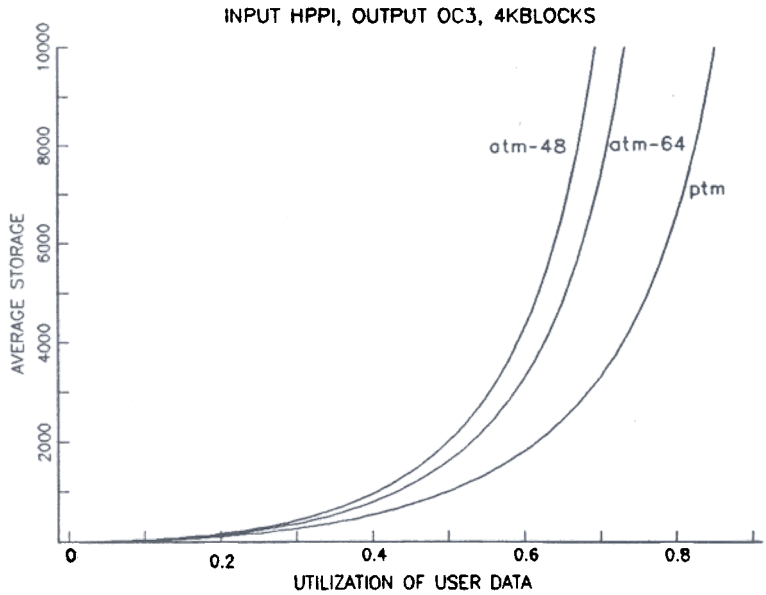
Results.

Again, we present a small set of representative results. We first pick an environment where the access links are composed of several FDDI rings (100 Mbps) and the output link is a OC3 SONET attachment (155Mbps). Packet sizes are 4Kbytes. We have plotted three results, ATM-64, ATM-48 and PTM-4K. The key observation is that while PTM offers higher delay at low utilization, it offers substantially lower delay at higher utilization. The crossover point is between 50

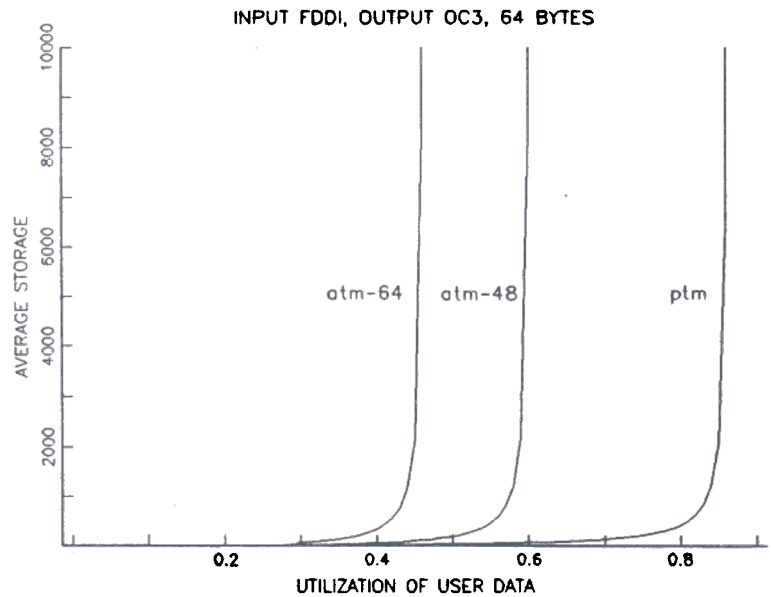
and 60 percent for this example. For low utilization, a reasonable argument can be made that delays do not matter since they are so low (in the absolute sense) for both PTM and ATM.



The second curve represents an environment where HPPI interfaces (800Mbps) interfaces act as sources of the traffic instead of FDDI rings. The clustering of cells becomes more pronounced and the advantage of ATM even at low utilization disappears. The crossover point is now close to zero utilization.



Finally, we present a somewhat unfair (for ATM) curve where all packets are exactly 64 bytes in length. The early saturation of ATM (caused by the packet integrality restriction) becomes the major factor.



Adaptation layer processing

The basic claim made in this section is fairly obvious. Since PTM has a much larger maximum packet size than the ATM cell size, it requires substantially lower segmentation and reassembly effort and consequently lower adaptation layer processing. The main contribution of this section is to quantify this difference and to thereby establish a significant drawback of ATM.

Analysis

Our focus here is on the receiver. This receives streams of ATM cells or PTM packets from the network and has the job of recreating a user packet stream. To estimate the burden of adaptation layer processing, we assume that a packet or cell takes X instructions worth of adaptation layer processing (excluding the processing unrelated to the adaptation layer). Based on our experience in design of transport layer protocols, we estimate X to be about 100 for typical cases. Additionally, we assume that the rate (bits/sec.) or user traffic is given by B and the traffic is drawn from one of the previously discussed offered traffic distributions. We have the following expression for the average number of instructions per second required at the receiver for an ATM cell size of A.

$$Avg.\ instructions\ per\ second = \sum_{u} X \times ceiling(\frac{p}{A}) Pr(random\ byte\ is\ from\ user\ data\ of\ size\ u)$$

A similar expression is obtained for PTM, replacing A by P. The distribution is exactly the same as required in the previous section.

2.6.2 Results

We have plotted two representative results below. In both cases, the speed of the user stream, B, is set to 16 Mbps and the value of X is 100. The first curve is from the simulated image application and involves large packets.

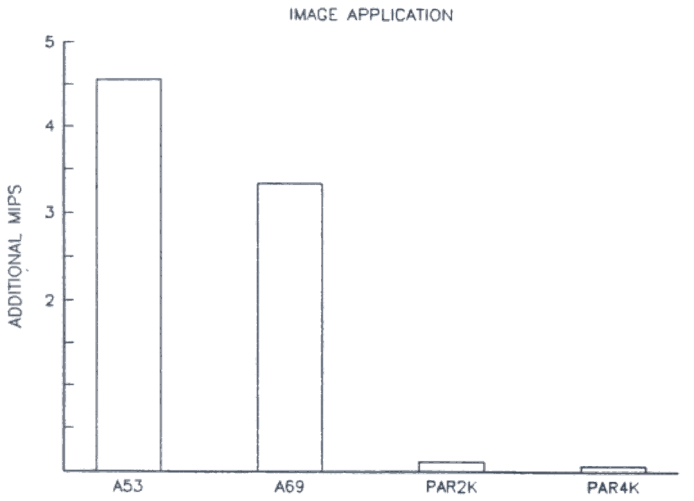
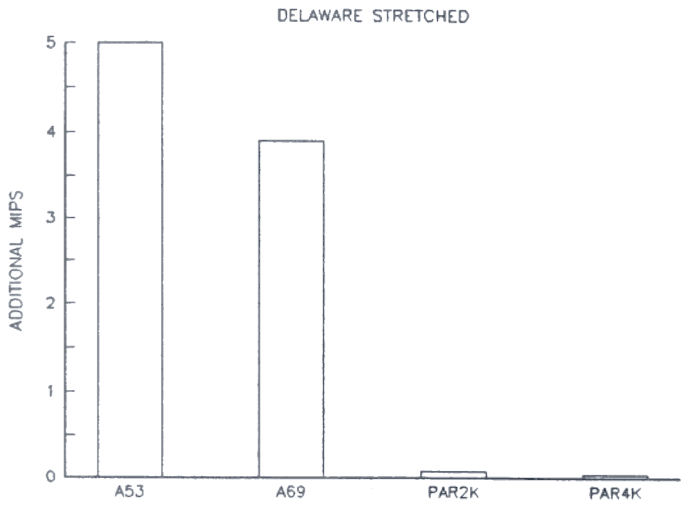


Image application

The second curve is from the Delaware stretched distribution and involves a much wider range of packet sizes.

2.6.3 Conclusions

We conclude that ATM may require custom VLSI for adaptation layer processing while PTM can make do with software processing handled by a relatively cheap microprocessor for network speeds of interest. This has major implications on the design and development cost of the adaptors (cheap microprocessors are substantially cheaper than custom VLSI). In addition, the reliance on custom VLSI limits the flexibility of the processing. (It may desirable to tailor the segmentation/reassembly function to the specific application. Will the ATM custom chip have some level of programmability? Clearly this will add cost and may also reduce speed. In addition to the limit on flexibility, the reliance on VLSI places a limit on the number of simultaneous connections that can be processed. In particular, an adaptor which has the responsibility of reassembling ATM cells into user packets may simultaneously have to reassemble a large number of packets. For example, a LAN adaptor may have thousands of active connections and may have to reassemble thousands of packets. Reliance on hardware limits the number of simultaneous reassemblies to the number of hardware instances of the adaptation function than can be placed on the adaptor.



Delaware

2.7 Statistical multiplexing

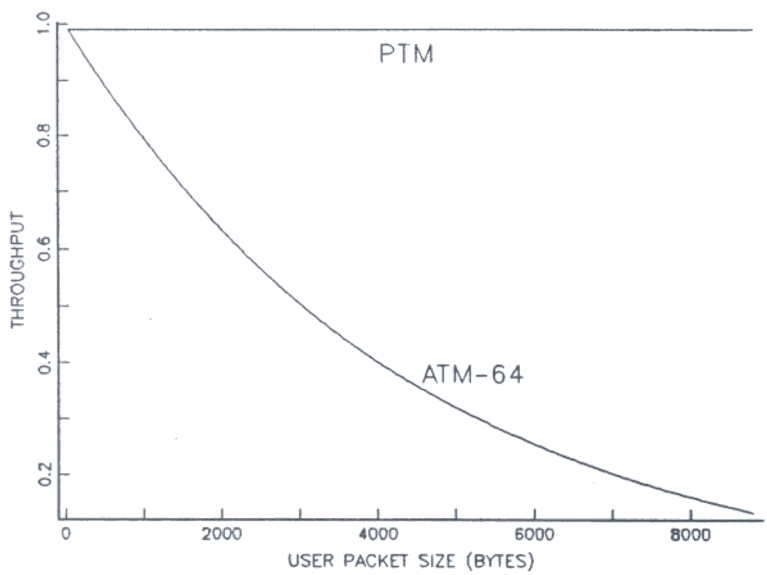
Statistical multiplexing is used in this section to denote the allocation of capacity to connection according to "average" requirements rather than peak requirements. In other words by exploiting the fact that a data connection is active only 20% percent of the time, we can pack in 5 times as many such connections over a link of a given capacity. This technique carries a risk- namely, the probability that for short periods more traffic is generated than can be carried by the network, resulting in packet loss. The claim is that by being somewhat conservative and by resorting to the law of large numbers it is possible to keep the probability of packet loss to small values. Clearly, as ATM is claimed to be a general transport mechanism, it must permit statistical multiplexing.

The main claim of this section is to demonstrate the pitfalls of statistical multiplexing in ATM. It is based on the observation that if packet discarding is necessary during temporary overloaded situations, it makes much more sense to discard units of information which are close to the natural unit of retransmission for the user. In other words, if the user understands packets and retransmits packets, it is inefficient to discard cells without any reference to packets. Even a relatively small cell loss probability is effectively multiplied many times because of the fact that a single cell loss results in the loss of a much larger unit (packet). We call this effect as tha AVALANCHE effect.

The current status of the proposed error recovery scheme of ATM (part of adaptation layer) calls for error recovery of lost cells using a retransmission at the message level. This means that a single cell loss (44 bytes of data) will result in a retransmission of the entire original packet. Since packets can be fairly large (8Kbytes packet is segmented into around 180 cells) this effectively increases the network loss probability for packets by at least 2 orders of magnitude. This is because any cell loss within the packet results in a retransmission of the entire packet. Another drawback of this scheme is that the rest of the cells of the message are still occupying space in the buffer and are even forwarded to the destination despite their expected retransmission. This inherently consumes more bandwidth in the following intermediate nodes making them more congested and increases their cell loss probability. This overall phenomena is the avalanche effect.

In contrast, the PARIS network keeps the error recovery units to be the same as the transmission units. This is clearly practical since PARIS accommodates relatively large packets (2-8KBytes depending on the implementation). In this case the packet is always fully discarded as a single unit.

On average, the loss probability of the network can be designed (by increasing the number of buffers at intermediate nodes) to compensate for this phenomena. However, when we are dealing with bursty processes like datagram, compressed video etc, there might be short period of times in which the network is over-utilized. The network may also be over-utilized because of timing mismatches in the bandwidth reservation process.



In the figure above we demonstrate that even under slight over-utilization like 1%, the avalanche effect may cause a dramatic decrease in the effective throughput for these traffic classes that require

error recovery. Intuitively, if packets are of the order of hundreds of cells even 1% loss probability causes almost every packet to be lost and retransmitted over and over again.

The negative effect of decoupling the error recovery mechanism from the lower layers is increased dramatically if more sophisticated congestion control mechanisms are used. Such mechanisms are employed in order to better exploit the statistical nature of bursty processes to increase the network efficiency. These proposals exploit the fact that certain traffic types employ end-to-end error recovery and thus can be sent at higher risk of being lost in the network.

In [ECK89], [BCS90] a source marking or coloring scheme is proposed in order to improve the statistical multiplexing efficiency of the network for bursty data sources. In this scheme, excessive load which is beyond the original bandwidth reservation of this connection is not simply dropped at the input. The rational is that the path might be in practice lightly utilized since many connections (just as bursty as this one) can be idle. Moreover, extensive amount of capacity might not be reserved at all over the path. The idea is that instead of dropping the packet at the source which is a sure loss, the source can take a calculated risk and send this extra packets by marking them as such. In order to guarantee fairness and acceptable service for other connections these marked packets do not have the same "right of way" as the unmarked packets. The intermediate nodes discards these packets first at time of congestion (for example, by using a lower discard threshold) such that the impact of the marked traffic on the well behaved bandwidth reserved traffic is minimized. If the marked traffic makes it way through the network this is clearly an advantage since some unused bandwidth was saved. If the marked packets are lost no harm was caused since they were considered as excessive anyway and since the other choice is to drop them at the source.

PARIS exploits such a congestion control mechanism in order to better support bursty connections where the notion of connection's average rate varies in time. The marked traffic is used in order to fill the gap between the actual increase of the connection rate and the completion time of bandwidth reservation (which requires a setup period delay). Unfortunately, as ATM is currently defined, the use of such a scheme can cause severe performance problems. First, we have already demonstrated that even a small amount of over-utilization can degrade dramatically the network performance. So the potential gain of such a scheme is quite small. Second, the fact that the ATM congestion control mechanism (where the marking is done) is architected to be bellow the adaptation layer (packet segmentation and error recovery) implies that it might be the case that cells of

the same packets will be marked differently. Since loss probability of marked cells can be considerably higher than that of normal cells the existence of even single marked cell increases the overall packet loss probability dramatically. This in fact implies an inherent impact of the marked cells on the normal cells which makes such proposals inadequate under the current ATM architecture.

3.0 LABEL ROUTING VS. ANR

In this section, we present and discuss various issues that arise because of the restriction to small fixed sized cells in ATM and compare it to PTM.

The ATM routing mechanism is based on label (VCI) swapping. The implementation at an intermediate node in the network requires the following operations to be performed for each incoming cell at the port adapter. A table look-up operation is required to determine the outgoing adapter as well as the new label needed for the next hop. After this, the actual header swapping is carried out for the cell and finally the cell is routed to the appropriate output port adapter. Apart from these "steady state" operations for routing, the VCI tables need to be updated for each call setup or termination.

In contrast to this, ANR method does not require any per connection table for steady state routing. The full routing information is part of the contents of the packet header. The first ID in the ANR field directs the message to the appropriate output port. The only operation needed at the output port is the stripping of this ID.

The potential disadvantage of the ANR routing compared to label swapping is the header overhead. The length of the initial ANR header is of the size of the call path. However, most networks are designed to have short end-to-end paths (4-5 hops) and the average length of the ANR is only half of the path length. This means that on average ANR headers are comparable in size or only slightly longer than label swapping headers. On the other hand, there are many disadvantage with label swapping as detailed below.

3.1 Nodal hardware/software complexity

For steady state switching, the nodal hardware needs to built carefully to ensure that the table look up, swapping and routing can be performed in "real" time. For a OC48 (2.4 Gbits/sec) this translates to about 170 ns to process each cell.

In ATM, the VCI field is defined to be of a length of 2 bytes. This limits the number of connections that can be supported over a given link to be below $2^{16} = 65,536$. If we consider a future OC48 line and 32KBPS voice connection the number of different voice connections is 75,000. With the addition of low speed data connection and possible future growth of the trunks speeds this

limitation seems to be a future obstacle. One way to solve it with no change to the cell structure is by using the concept of VPI (virtual path identifiers) i.e. multiplexing multiple connections that share the same end-to-end path with the same two bytes ID. This concept will add complexity and constrains to the routing and control mechanisms of the ATM network. Another alternative is to allow for a 3 bytes VCI Ids. However, given the extremely small size of the cell, this implies a significant drop in the network utilization efficiency.

In contrast to the above, the is obvious advantage of ANR is the simplicity of intermediate node processing. No large tables are needed and the swapping operation is eliminated. Since no tables updates are needed, the control information exchange between the line adaptors and the nodal control point for route establishment is minimized. This off-loads from the control point software a significant part of the call processing overhead. In addition, with the ANR method there is no limitation to the number of different connection that can be supported over any speed trunk.

3.2 Connection setup and termination

We have already described the fact that no routing table are needed for the call establishment. If the call information (such as bandwidth, class of service and burstiness) needs to be recorded by the nodal processor this can be done "on the fly" using the copy capability of the ANR mechanism as shown in figure.

The copy function allows for selected packets to be copied by the nodal processor in addition to being switched over the path specified in the ANR header. This permits a single setup/takedown message to be used in parallel in order to update the call information at all nodes along the call path.

If the ATM networks supports only VCI based connections the call setup process has to be hop-by-hop. The setup/take-down messages will be forwarded as single hop messages with potential software processing at each hop. This sequential processing at each hop can considerably slow down the setup/take-down mechanism.

Another important issue *transparent route switching*. This service makes network internal failures transparent to the end user by rerouting calls very fast after such events. In ATM, the need to setup VCI IDs along the new calculated path and their release along the old path makes this hard to handle. In the PARIS network this process is significantly facilitated since no such actions have to be taken.

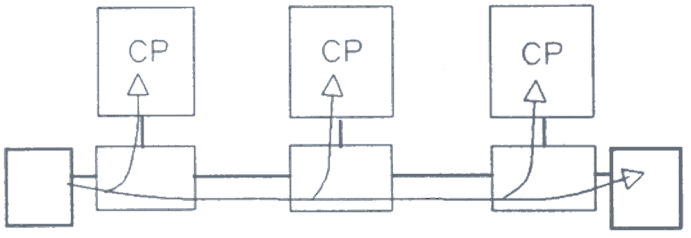


Figure 1 Selective copy mechanism.

3.3 Datagram support

An important feature of the ANR scheme is that there is no need for a table setup before packets can flow over the connection between the end-points. This enables an efficient support of datagram. Datagrams can flow with no nodal software involvement (except for the initial route computation or a single look-up operation at the source) through the switching hardware directly to the appropriate end-points. Using this features PARIS can support datagram type traffic by either sending them directly to the destination from the source (and maintaining a cash of routes at this source) or by sending them to a datagram server which maintains a larger table of routes to potential destinations; and from there switched directly to the destination. This is much more simple and efficient than the way that SMDS has to support datagram services through multiple servers. These multiple servers must be connected by predefined set of sessions which form a virtual network of servers on top of the ATM network. This will cause either a maintenance of a large amount of such connections (the extreme case is a fully connected graph) or to relatively long paths through the servers' logical networks (the extreme case is a tree based structure).

4.0 SUMMARY AND FUTURE WORK

In this paper we have attempted to present our early concerns regarding ATM and brought them to a sharper focus by comparing it to PARIS.

We have used quantitative evidence to show that ATM cell transport is not optimized for data. Variable length packets have significantly better performance characteristics. In regard to label swapping our conclusions are somewhat preliminary. A thorough evaluation is difficult without a detailed side-by-side comparison based on operational prototypes of the technologies involved. This is precisely one of the goals in the Aurora project that is just getting underway. Aurora is a Gbit/sec network testbed planned under the umbrella of the Corporation for National Research Initiatives (CNRI). The participants are Bellcore, IBM Research, MIT and University of Pennsylvania. The plan to install and operate two 4 node testbeds located at the sites of the participants. The first testbed will be built by Bellcore based the Sunshine ATM switches and second will be built by IBM, based on the PARIS architecture. These testbeds are planned to installed in '91. This effort will allow us to understand and compare in detail both the issues related to advantages/disadvantages of the two approaches as well as the inter-operability issues.

Special acknowledgements We would like acknowledge Khosorow Sohraby and Kip Potter for providing valuable input during the course of this work.

5.0 REFERENCES

[CG88] I. Cidon and S. Gopal, "PARIS: An Approach to Integrated High-Speed Private Networks," International Journal of Digital and Analog Cabled Systems, Vol. 1, No. 2, pp. 77-86, April-June 1988.

[BCS90] Bala, K., Cidon, I., Sohraby, K. "Congestion Control for High Speed Packet Switched Networks" INFOCOM 1990, To Appear.

[BGGJP85] A. E. Baratz, J. P. Gray, P. E. Green Jr., J. M. Jaffe and D. P. Pozefsky, "SNA networks of small systems," IEEE Journal on Selected Areas in Com., Vol. SAC-3, No. 3, pp. 416-426, May 1985.

[IEEE88] IEEE 802.6, "Proposed Standard: Distributed Queue Dual Bus Metropolitan Area Network", 15 November 1988.

[SOH89] K. Sohraby, "Delay Analysis of a Single Server Queue with Poisson Cluster Arrival Process arising in ATM Networks" Globecom '89 Dallas, Vol. of 3, pp.0611-0616, November 27-30, 1989.

[SMDS89] Bellcore, "Generic System Requirements in Support of Switched Multi-megabit Data Service", TA-TSY-000772, Issue 3, October 1989.

[ATMA90] CCITT SG XVIII, "Draft Recommendation I.121: Broadband Aspects of ISDN", Geneva, January 1990.

[ATMB90] CCITT SG XVIII, "Draft Recommendation I.150: B-ISDN ATM Functional Characteristics", Geneva, January 1990.

[MIT86] D. Feldmeier "Traffic Measurements on a Token Ring Network," Computer Networking Symposium, Washington, DC p 236-243, November 17-18, 1986

[UD87] P. Amer,R. Kumar, R. Kao, J. Phillips, and L. Cassel "Local Area Broadcast Network Measurement: Traffic Characterization," IEEE Computer Society International Conference 32nd, Piscataway, NJ,1987.

[ECK89] S. Eckberg, D. Luan, D. Lucantoni, :"A Congestion Control Stategy for Broadband Packet Networks - Characterizing the Throughput-burstiness Filter," International Teletraffic Congress Specialist Seminar, Adelaide 1989.

[BER87] R. Gusella, "AD-A196100 /2/XAB Analysis of Diskless Workstation Traffic on An Ethernet," Technical rept. Aug 7 1984-Aug 6 1987, pp.27.