Rationalizing Key Design Decisions in the ATM User Plane

Daniel B. Grossman
Motorola, Inc.
dan.grossman@motorola.com

Abstract

Any technology requires some number of key design decisions. In the case of the ATM user plane, the choices to use fixed length, 53 byte cells and virtual connections were unorthodox from the perspective of many in the networking research community. This paper attempts a technical justification for those design decisions.

1 Introduction

Technological superiority is not the sole determinant of market success in technology wars; witness, for example, the results of the Betamax vs VHS wars. ATM, which was once hoped to be the underlying substrate for next generation multiservice networks, never achieved its potential. Its proponents permitted it to be drawn into a technology war with IP, and in the end, it was badly battered in the marketplace. Its opponents were able to portray it as complex and inefficient, and this image did much to undermine it.

ATM's largest vulnerability in the war of perception was in three related design decisions: its use of fixed sized protocol data units (PDUs), or cells, the 48-octet cell payload size, and its stateful, or connection-oriented nature. These were very different from design decisions made in IP, and thus unorthodox from the perspective of many in the networking research community. Some of the overhead that resulted from these decisions quickly attracted a derisive tag: the cell tax. This stuck, and contributed significantly to ATM's image problem. This paper presents a technical defense for these design decisions.

2 Fixed Size Cells

Fixed sized PDUs are the most singular design decision in the ATM technology. Much of the benefit of ATM, and many other design decisions, flowed from the fixed size cell.

2.1 Deterministic service times

Fixed length PDUs have deterministic service times. Deterministic service times reduce average waiting times in congested systems. As an approximation, recall that an M/D/1 queuing system has half the average waiting time of an M/M/1 queueing system [1]. This means that cell-based networks can operate with higher link utilization than variable packet based networks while maintaining acceptable delay and loss. Anecdotally, the Internet backbone is engineered to 30% average loading, while ATM networks can be engineered to 80% average loading [2]. Deterministic service also leads to tighter upper bounds on delay for real-time traffic than non-deterministic service times. This is especially useful when real-time and non-real time traffic shares the same physical links, for low-rate links, and in the presence of constant rate real-time traffic having dissimilar rates.

Queueing systems with deterministic service times are easier to analyse than those with general service times. This property can be used to simplify the design and reduce the runtime complexity of admission control, scheduling and other networking algorithms.

2.2 Granularity mismatch

Variable length packet forwarders can only forward integral packets, yet must account for rate and other functions of time in units of bytes. This inconsistency can perturb network dynamics and add complexity. In the ATM forwarding path, the cell is only unit of interest, and no such inconsistency exists.

For example, policers in variable length packet networks accumulate credits in units of bytes, while admit/mark/ drop decisions must occur on packet boundaries. Thus, if a packet arrives when the number of available credits is positive and non-zero but less than the packet length, the policer must either mark or drop the whole packet, or 'borrow' credits against future packets. Each of these alternatives discriminates against larger or smaller packets (respectively), and over or under polices (respectively) [3]. Since policers in cell-based networks make admit/mark/drop decisions and accumulate credits on units of one cell, they more precisely police to traffic contracts. Thus, admission control decisions can be more aggressive without violating service guarantees.

Similarly, schedulers in variable length packet forwarder can decide which of several queues to service next next only on a per-packet basis. When the scheduling policy is fairness (or weighted fairness), a simple scheduling algorithm, such as (weighted) round robin or (weighted) fair queueing is biased against sources that send small packets. More sophisticated algorithms, such as deficit (weighted) round robin, are required to correct this bias, but they add complexity and discriminate to a lesser extent against sources that send larger packets [4]. In networks with fixed length PDUs, the unit of scheduling is the size of the PDU, and therefore no such discrimination can occur: simpler schedulers are fair.

For real time flows, it is desirable to service PDUs as close as possible to an ideal departure time. Various kinds of calendar scheduling schemes have been used to do this. Fixed sized PDUs greatly simplify the calendar scheduling problem, since the size of the PDU need not be taken into account in determining whether it can be scheduled before a more urgent one.

2.3 Hardware Implementation

Queueing systems, buffer managers, interconnects and other datapath elements can be greatly simplified by fixing the size of the PDUs they carry. Complexity metrics such as interconnect width, memory sizing, number of control lines, and amount of handshaking can be reduced by fixing the length of PDUs. These metrics can affect die size, clock speed and pin counts. For example, a fixed sized PDU allows pipeline stages to be made synchronous, permits parallel processing without having to perform internal buffering to avoid misordering, and requires only a single size buffer pool, without need for fragmentation. Deterministic service times remove blocking, and facilitate cycle count budgeting. These advantages are so significant that many Ethernet switches and IP routers use fixed size data units in their fabrics.

3 53 Octets

The 48-octet payload and 5-octet header of the ATM cell were a notorious political compromise between the needs of speech and data transmission. At the time that the cell size needed to be set, the French Administration planned a telephony transport network with budgeted packetization delays of 4 ms, or 32 octets per cell with 64 kbit/s PCM; this was small enough to avoid the need for echo cancellation on calls within continental France. Opinion at the time was that optimal cell size for data transport would be between 64 and 256 octets. The compromise, debated over several meetings of the former CCITT SG 13 [5], is not quite optimal for either telephony or data, but may be globally optimal for the mix of traffic that would be carried in a multiservice network.

- For PCM telephony, a full cell has a packetization delay of 6 ms, which is usually within the delay budgets of modern Voice-over-ATM systems, even those designed for transcontinental calls.
- Measured packet size distributions in the Internet have a sharp peak at 40 octets, which is the size of an IPv4 datagram containing a TCP SYN, FIN or ACK, with neither IP nor TCP options. Such a packet fits exactly into a single cell when AAL5 is used. Unfortunately, LLC/SNAP encapsulation is frequently used instead of the so-called null encapsulation. It adds 6 octets of overhead, and therefore forces TCP control packets to cross a cell boundary.
- Two MPEG2 transport streams packets exactly fit into eight cells when AAL5 is used.

In addition, most switch implementations require some internal overhead to be prepended to each PDU to form an internal forwarding unit. Switch hardware usually operates in units of bytes that are powers of 2; thus a 64 bytes internal forwarding unit (including 16 bytes of overhead) can be extremely convenient.

The "cell tax" slur arises from the five octets of overhead required for 48 octets of payload. Recollect that all packet-based networks have header overhead. ATM cell header overhead is 9.5%. PCM telephony systems that can accommodate a 6 ms packetization time suffer only this overhead. MPEG2 transport over ATM suffers an additional 3.2% overhead due to the AAL5 trailer, in the default mode of operation (two MPEG2 TS packets in an AAL5 PDU). For IP transport over AAL5, including LLC/SNAP encapsulation, overhead as measured in the Internet is 20 percent [6]. Without LLC/SNAP encapsulation, it would be 15%, leading one to speculate that if header overhead were truly meaningful to network operators, LLC/SNAP encapsulation would have been eliminated. More important, operators can more than compensate for ATM header overhead by more aggressive traffic engineering, which is made possible by deterministic queueing.

4 Virtual Connections

Packet-based networking systems may be characterized as being either connection oriented or connectionless. Selection of one of these design approaches has broad implications on the network architecture and its implementations. This has been a deeply controversial subject since the early 1970s. For the environment for which ATM was intended, connections and their attendant state information have a number of advantages, and some disadvantages.

Much processing is done once in connection-oriented systems during connection establishment, but done for each packet in connectionless systems. This makes connection-oriented systems better optimized for long-lived communications, as would be expected for telephony, video delivery, and similar applications. Connectionless systems are better optimized for short lived communications such as DNS lookups. A majority of packets in the Internet are associated with long-lived flows, even though the majority of flows are short lived [7]. Attempts by the ATM Forum to create a native connectionless mode to complement the connection oriented mode were unfortunately not successful. On the other hand, one could imagine that applications might have evolved differently if the underlying network encouraged design for long-lived flows.

Per-packet forwarding decisions in ATM nodes are can be implemented with simple table lookups or binary content-addressable memories, over the VPI and/or VCI fields, with O(1) computational complexity. The size of these tables can scale with the number of VP and/or VC links expected to cross each interface. IP requires longest match lookup over the IPv4 or IPv6 destination address fields. To the author's knowledge, the least computationally complex longest match algorithm has lookup complexity of O(LogN) and update complexity of O(kLogkN) [8]. Ternary CAM devices help, but are complex and presently expensive. Forwarding tables scaled to global Internet presently need tens of thousands of entries.

State information is required by policers, shapers, markers, schedulers, per-connection queues, flow control and admission control, which are needed to support services with QoS guarantees. In an ATM network, this state information is inherently associated with a virtual connection link through the VPI/VCI. In connectionless networks, packet classification is necessary to associate a packet with its state information. Classifiers have between O(logN) and O(N) complexity, and mechanisms and policies for configuring classifiers (especially in the core) are complex and likely to be imprecise.

State information can be used in conjunction with per-connection queues and scheduling algorithms to enforce network policies (e.g., fairness) among best effort connections. Connections that transmit at a rate exceeding their share of link bandwidth fill their queues, invoking discard policy. As a result, these connections do not interfere with better-behaved connections. Forwarders for connectionless networks may, of course, implement classification and per-flow queueing, at cost in complexity as noted above. Otherwise, misbehaved connections can at best be dealt with in a statistical fashion, using an active queue management algorithm such as RED. While there is some evidence that such algorithms improve the stability of the Internet [9], there is also evidence that they are not effective [10,11], that they are hypersensitive to configurable parameters [12], and that they do not address non-responsive flows. Per-flow queueing and longest queue discard are a far better solution.

Connection-oriented networks do not suffer from many of the security vulnerabilities that are often exploited in connectionless networks; for example, attacks that depend on address spoofing, such as the SYN attack, are not possible in connection oriented networks. State information also makes tracing an attack back to its originating interface significantly easier. Further, state information is useful as a basis for security associations and audits, and simplifies security design and implementation. In addition, usage accounting, as required for usage-based billing, requires state information. In cases such as virtual private networks, IP-in-IP and IPSEC tunnels, with their considerable overhead, are used as a substitute for connection state.

5 Conclusions

ATM's key design objectives were to operate in a public network environment, to carry real time and non-real time traffic and to be optimized for hardware implementation in switches. A series of design decisions flowed from those objectives, including virtual connections and fixed size, 53-octet cells. From a technology perspective, it can be argued that those decisions withstood the test of time. Unfortunately, the arguments for these decisions are subtle, and the counter-arguments were simple and were made persuasive by repetition. It is not clear how much the "cell tax" slur contributed to limiting the success of ATM. There were other issues, such as the experience curve resulting from a large previously installed base of Ethernet, slow deployment of capabilities to simplify configuration (especially SVCs), lack of ATM-aware applications software, delays in critical standards, and numerous business decisions made by some industry players. Nonetheless, market perception was an important factor, both directly and in its effects these other issues.

Perhaps the most important lesson that can be drawn from this paper is that proponents of novel networking technologies need to justify their design trade-offs in a way that will be understood by the market, and to reinforce their justification. They must also be prepared respond forcefully and credibly when these trade-offs are attacked. ATM's proponents did neither.

References

[1] Kleinrock, Leonard Queueing Systems vol. 1 (New York, John Wiley and Sons, 1975) p. 191

[2] Private conversations

[3] Bernet, Y; Blake, S. Grossman, D; Smith, A "An Informal Management Model for Diffserv Routers" RFC 3290 (April, 2002)

[4] Shreedhar, M.; Varghese, G. "Efficient Fair Queueing using Deficit Round Robin" *ACM Computer Communication Review*, vol. 25, no. 4, pp. 231–242, Oct. 1995.

[5] Reports of the meetings of CCITT SG 13, 1988

[6] Thompson, K.; Miller, G.J.; Wilder, R. "Wide-area Internet traffic patterns and characteristics" *IEEE Network*, Volume: 11 Issue: 6 , Nov.-Dec. 1997 pp. 10 -23

[7] ibid.

[8] Ruiz-Sanchez, M.A.; Biersack, E.W.; Dabbous, W. , "Survey and taxonomy of IP address lookup algorithms" *IEEE Network* , Volume: 15 Issue: 2 , March-April 2001 pp. 8 -23

[9] Braden, B. et al, "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309 (April 1998)

[10] May, M.; Bolot, J.; Diot, C.; Lyles, B., "Reasons not to deploy RED", technical report, June 1999.

[11] Christiansen, M; Jeffay, K; Ott, D; Smith, F.D. "Tuning RED for Web Traffic", *Proc. ACM SIGCOMM*, August 2000

[12] Firoiu, V.; Borden, M., "A Study of Active Queue Management for Congestion Control", *Proc. IEEE Infocom 2000*