MEETING THE CHALLENGE: CONGESTION AND FLOW CONTROL STRATEGIES FOR BROADBAND INFORMATION TRANSPORT

A. E. Eckberg, Jr.
D. T. Luan
D. M. Lucantoni

AT&T Bell Laboratories
Holmdel, NJ 07733 USA

ABSTRACT

Determining approaches to congestion and flow control, especially real-time components in an overall strategy, is recognized as one of the fundamental challenges facing broadband "packet-based" information transport, as, e.g., in the case of B-ISDN/ATM. In this paper we summarize basic issues underlying this subject, and describe a particular approach to achieving a broadband congestion, flow, and error control architecture, based on a "core" congestion control strategy that we term *Bandwidth Management*. The modular and layered nature of this control architecture is described, and shown to lend itself to a structured approach to characterizing the control architecture performance.

1. INTRODUCTION AND SUMMARY

Broadband ISDN (B-ISDN) architectures, built on a technology base of information transport via the Asynchronous Transfer Mode (ATM) protocol[1] [2], offer a means for the efficient integrated transport for a broad spectrum of services, from voice and low-speed data to high-speed data and video telephony. While well-thought-out engineering and administration rules will likely result in excellent end-to-end performance in such networks most of the time, a fundamental challenge in the design of robust B-ISDN architectures is the structuring of a comprehensive strategy for congestion and flow controls, which will be essential during those (hopefully rare) instances of resource overloads due to unforeseen traffic focusing and peaking, as well as during network component failures. Unfortunately, the multi-megabit-per-second speeds that will be responsible for the high performance/capacity of B-ISDN/ATM networks will also severely limit the number of implementable and stable congestion control strategies.

In this paper we address a number of basic issues in the area of congestion and flow control strategies suitable for B-ISDN/ATM, and suggest a specific B-ISDN/ATM congestion/flow/error control architecture. We begin, in Section 2, by summarizing goals, issues, and several guiding principles that underlie the selection of such a control architecture. Section 3 describes a *Bandwidth Management*-based congestion/flow/error control architecture that is motivated by some of the considerations laid out in Section 2. Then, Section 4 describes a modular approach to achieving a traffic/performance characterization of this control architecture.

2. UNDERLYING ASSUMPTIONS, MAJOR ISSUES, AND GUIDING PRINCIPLES

There are two, fundamentally different, approaches to determining control architectures for complex systems such as B-ISDN/ATM networks: an *inside-out* approach, in which each component in an overall architecture is looked at individually, without excessive attention given to component interworking, and an *outside-in* approach, in which a global, "system" view is taken of the entire system. It is the inside-out approach which typically must be taken during experimentation in technology advancement and component prototyping, in particular to determine component functionalities that are feasible to implement. However, at an appropriate point in technology development it becomes necessary to switch to an outside-in approach, so that the entire system can be characterized as a single entity, and so that system-wide strategies that may make global sense are not unnecessarily prevented by individual component limitations.

It is an outside-in approach which we attempt to take in this paper, focusing on issues and characteristics of entire B-ISDN/ATM information transport systems. Therefore, in this section we list some of the assumptions we are making concerning the characteristics of B-ISDN/ATM networks and the information movement and management environments they will be supporting. These assumptions then lead to a number of high-level objectives and guiding principles for a B-ISDN/ATM congestion/flow/error control architecture.

2.1 Assumptions

We first clarify the framework underlying the control architecture by listing several key assumptions:

i. We restrict attention to B-ISDN/ATM networks, and the *connection-oriented* information transport service they will provide. In generic terms, we think of such networks as consisting of network "nodes" and a network "edge," where the latter provides the B-ISDN/ATM interface between the "network" and ATM end-devices. We specifically include these ATM end-devices within the definition of the "B-ISDN/ATM information transport system," because there is a fundamental role such end-devices will play in an overall B-ISDN/ATM control architecture, particularly in the area of flow and error controls. These ATM end-devices may be end-terminals that handle the ATM protocol themselves, or terminal adaptors (TAs) allowing end-terminals associated with a variety of non-ATM services and native protocols to adapt to the common ATM transport.

ii. For the scope of this paper, we further assume that the network itself is a single entity, under the administrative control of a single "owner." Thus, we will not concern ourselves with control issues associated with interconnected networks.

iii. We specifically assume that the B-ISDN/ATM network will provide the integrated transport for a variety of different services, only some of which are data services. Moreover, there is a great uncertainty associated with many of these services and applications, particularly in regards to their traffic characteristics and their performance needs (e.g.,

CH2682-3/89/0000-1769 $1.00 © 1989 IEEE

low-delay, or low-loss). This uncertainty impacts on the desirable characteristics of a control architecture, since it points to the need for considerable robustness and flexibility in accommodating diverse services, some of which may not even be conceived of until they have been stimulated by the existence of B-ISDNs.

iv. We also point out that ATM end-devices cannot be assumed to be under the total control of the network, particularly when the network is a public, shared one. One implication of this observation is that the network cannot assume a "friendly community" of end-devices, and in particular, cannot assume that individual end-device actions will be in the best interest of the network and other end-devices. Even if end-device algorithms were to be completely standardized, there is still the possibility that one or more end-devices would not conform to the standards in certain critical, and difficult-to-detect, areas.

2.2 Major Issues in the Design of Broadband Control Architectures

A basic characteristic of packet-based information transport, such as with B-ISDN/ATM, is the sharing of critical network resources (memory, processor real-time, transmission bandwidth) between large numbers of information flows. Controls are needed to protect such shared resources from depletion and to allocate them fairly, and the difficulty in achieving effective controls generally increases with increasing transport speeds. At broadband speeds three factors become particularly notable:

i. The level of control algorithm complexity that can be implemented in individual devices is limited due to the costs associated with specialized high-speed hardware.

ii. The effectiveness of sharing explicit control and status information between the various ATM cell-handling devices along a transport path, to trigger continuous, real-time control actions, is limited by large propagation-delay-to-cell-queue-time-constant ratios[1].

iii. Transport integrity (without reliance on recovery mechanisms) becomes a difficult goal in many settings, since large and expensive buffering memories may be needed to eliminate ATM cell losses in the network, while if such large buffering memories are used, this adds significantly to end-to-end tail delays.

This last factor implies that ATM cell loss during broadband transport is a phenomenon that may be difficult to "design away"[2], and thus it becomes important that error control strategies be derived jointly with congestion and flow control strategies.

2.3 Guiding Principles in the Design of a Broadband Control Architecture

The above-mentioned assumptions and issues motivate the following principles to guide the design of broadband congestion/flow/error control architectures:

i. The control architecture should be *flexible*, capable of accommodating the characteristics and requirements of yet-to-be-determined mixes of services and applications.

ii. The control architecture should be *robust*, in that its operation and resulting performance should not depend strongly on assumptions concerning the traffic characteristics of services, the cooperativeness of end-devices, or even basic transport characteristics of the network (actual buffering capacities, propagation delays, etc.). In particular, we feel that systems of controls that are tightly-coupled between the network and end-devices tend to be intrinsically nonrobust, and that therefore if loosely-coupled, distributed controls can be designed, such controls may exhibit considerably more robustness.

iii. The control architecture should be based on *as simple as possible* control actions, in order to increase the chances of implementability at broadband speeds, and to result in an intuitively understandable overall system.

iv. The control architecture should be separable into three distinct domains, each with its own set of objectives:
   a. a set of network congestion controls, allowing the network to protect its shared resources and to allocate them "fairly" in accordance with resource-demand agreements between the network and end-devices; moreover, this protection and fair allocation should not depend on end-device actions.
   b. flow control actions (e.g., protocol window reduction and/or "traffic-shaping") taken by, and under the control of, the end-devices; such actions should be motivated by end-device self-interest, but should result in actions for the common good.
   c. error control actions, such as forward error correction[3], retransmission, or source coding to allow cell-dropping[4] [5] [6], that are synergistic to the congestion and flow controls.

v. The control architecture should be "layered," with simple, real-time, distributed "core" actions, overlaid by call-level actions acting over longer time scales.

3. CONGESTION, FLOW, AND ERROR CONTROL STRATEGIES BASED ON BANDWIDTH MANAGEMENT

This section provides a motivation and description of congestion, flow, and error control strategies based on *Bandwidth Management* (BWM). We begin with a summary of the BWM "core" network congestion controls themselves, then describe call-level controls needed as an overlay to the BWM "core," and finally observe how flow and error controls implemented in end devices or adaptors could be structured to complement the network congestion controls. It can be noted that the overall control framework described in this section is similar to ideas in other recent papers[7] [8] [9]. Moreover, the BWM core controls have been implemented in wideband packet technology devices[10].

3.1 The BWM Core Network Congestion Control

BWM is motivated by three principle control objectives: (i) *protection of shared network resources*, (ii) *fairness*, and (iii) *robustness;* and by two principle implementation objectives: (i) *simplicity* and (ii) *flexibility*. These five objectives reflect the end goal of achieving a manageable control strategy with good performance. The BWM approach to achieving resource protection is to rely on congestion relief through cell discard; and fairness is achieved with BWM by making the cell discard process *selective*, based on traffic agreements and real-time traffic monitoring. The simplicity objective is approached through the primary role played by selective cell discard (an intrinsically simple

1. For example, the transmission time for a 53-octet ATM cell is approximately 2.8 microsec. at 150 MBPS, and the resulting time-constant in a cell queue will be on the order of 10's of microsec. for moderate utilizations, while the end-to-end signal propagation delay through fiber for a 1000 kilometer connection is approximately 5 millisec.

2. In fact, *selective packet discard* can form the basis of a robust congestion control.

operation), and through a *distributed control structure* that involves local control actions, and no real-time coordination between network components, or between the network and end devices. Flexibility is achieved through the explicit role of traffic agreements and monitoring. Robustness is approached through non-real-time-intensive couplings between the BWM "core controls" and "call-level" controls.

The BWM core controls are applied to the cell flow of a virtual circuit (VC), and are based on a set of traffic parameters, established at VC setup for each of the VC end devices, defining the characteristics of traffic from that end device that the network is willing to support. These traffic parameters are directly related to a method for monitoring traffic in real-time, e.g., the "leaky-bucket" monitoring algorithm[11] [12], and establish a traffic throughput level and degree of burstiness that the network must transport with a high degree of integrity. The end device may exceed this traffic agreement, but then should not expect complete transport integrity. The core controls are as follows:

i. For the duration of the VC, traffic entering the network from each end device is monitored in real-time at the network edge, e.g., via the leaky-bucket algorithm; and through such real-time monitoring, traffic that falls within the agree-to traffic parameters is discriminated from the remaining traffic (the *excessive* traffic), if any. Cells corresponding to excessive traffic are then marked with a *violation tag* in their header, indicating that they should be transported, but at the risk of possible discard should these excessive-traffic cells encounter congestion along the VC route.

ii. Moreover, "pre-marking" of cells with a violation tag at an end device can also be permitted. The monitoring of traffic from the end device is not influenced by such "pre-marked" cells; rather, these will be ignored in the traffic monitoring, and made to retain their original violation tag. This allows an end device flexibility in designating those cells it may wish to have transported "at risk," but monitoring is still always performed at the network edge for protection. For example, voice and video end-devices could pre-mark non-essential cells.

iii. Violation-tagged cells are discarded whenever they encounter a moderate threshold at any ATM cell queue along the virtual circuit, so that the presence of excessive-traffic cells causes only negligible impact to the delays, and virtually no impact to the loss, of cells not carrying a violation tag.

It can be seen that the BWM core controls involve simple and distributed operations. In particular, no explicit passing of congestion status information between ATM cell-handling devices along the VC is involved; rather, simple local actions are taken in response to the sizes of local ATM cell queues. Once initialized by the establishment of traffic monitoring parameters, they are "free-running" controls. Yet, cells are discarded only when congestion relief is needed, thus "closing the loop" in real-time.

There are numerous variants of the leaky-bucket monitoring algorithm, but a typical one involves maintaining a counter with simple operations at successive cell arrivals on a VC as follows:

i. The counter value, X, is initialized to 0.

ii. At each new cell arrival, the counter value is first decremented as $X \leftarrow max(X-cT, 0)$, where c is a specified parameter, and T is the elapsed time since the last cell arrival (but, as indicated, X is not decremented below 0).

iii. The counter value X is then compared with a prescribed value M, and
   a. if $X \geq M$, the cell is violation-tagged and passed on;
   b. if $X < M$, the cell is passed untagged, and the counter is incremented as $X \leftarrow X+1$,

This algorithm allows a sustained, untagged, throughput rate of c cells per time unit, and a burstiness that is controllable (for a given value of c) through the selection of the value of M. One can envision the leaky-bucket algorithm as a *throughput-burstiness filter* applied to the cell flow entering the network from an end device on a VC. Note that the operation of the leaky-bucket is identical to that of a finite-capacity single server queue with a deterministic service time.

3.2 BWM Call-Level Controls To Overlay The BWM Core

As noted above, the BWM core controls are simple, distributed controls. They can be expected to provide the necessary resource protection and fairness when allowed to "free-run" over significant time intervals (i.e., "significant" with respect to characteristic time-constants of the network). However, they require initialization, and also may require minor "tuning," although not very frequently. This is the purpose of the "call-level" controls, which provide an additional "closed-loop" aspect to the overall network congestion control strategy.

In particular, at each VC setup there needs to be a negotiation to establish the leaky-bucket parameters for the traffic monitoring/tagging process. This entails determination by the network whether or not a requested set of traffic parameters can be supported, and thus whether the VC setup should be accepted, denied, or modified. This determination could depend in part on a quantitative prediction of the throughput and burstiness of the traffic that will be allowed to pass without a violation tag, if a particular set of traffic parameters is allowed. Thus, a traffic characterization of the throughput-burstiness filter is a building-block for the VC acceptance/denial/modification process.

Other operations that need to be performed at the call-level are as follows. Unless VC setup requests are handled very conservatively, there is always the chance that too much non-violation-tagged traffic will have been permitted, and that consequently network resources are being stressed. If such an event occurs, it will be necessary to invoke call-level controls to reduce the amount of non-violation-tagged traffic. Possible control actions include tear-down of selected VCs, renegotiation of traffic parameters on existing VCs, or simply denying all new VC setup requests for a fixed period of time.

3.3 End Device Flow And Error Controls To Complement The BWM Core

As was noted earlier, while with BWM the network-implemented congestion controls require no explicit coupling with end-device-implemented flow and error controls, the *strategies* underlying these controls need to be coordinated and consistent. This coordination could be very simple, or could be quite sophisticated, depending on the degree to which an end device wishes to exploit the transport characteristics that BWM provides. In particular, it may be noted that a VC through a BWM-controlled network has the appearance of a "statistical VC," with certain fixed characteristics such as "guaranteed" throughputs and burstiness that will be supported, and other statistical/dynamical characteristics (that change with network congestion) relating to how much excessive traffic can be sent effectively, albeit always at risk.

An end device may choose to implement flow controls which tend to keep traffic at levels where only a negligible fraction of its cells are discarded. This could be accomplished at the end device by "shaping" its traffic via a leaky-bucket algorithm and buffering, so that none of the cells will be violation-tagged by the network.

Alternatively, the end device may maintain a handshaking process with its peer device, to determine to what extent its excessive traffic is getting through. Note that if sending excessive traffic is to be attempted, this traffic must either be considered "expendable," or it must be protected via end-to-end error controls. When excessive traffic is being discarded with a high enough likelihood, the end device can then apply self-throttling or traffic-shaping methods to reduce the amount of excessive traffic, and thus reduce or eliminate the amount of cell discard. Such adaptive techniques for flow and error control could, e.g., be obtained via adaptive windowing when suitable protocol windows apply.

3.4 An Example Of BWM-Based Congestion/Flow/Error Controls: LAN-To-LAN Interconnect

We now sketch through a hypothetical networking application that illustrates the possible operation of the modules in the BWM-based control architecture. Consider a B-ISDN network with ATM transport, and suppose that several of the end devices, connected via a set of VCs, are Ethernet LAN (ELAN) terminal adaptors (TAs). We assume that each of these TAs supports a number of 10 MBPS ELANs, and that each TA has a 150 MBPS access link to the B-ISDN network. Traffic from one ELAN to another ELAN not served by the same TA will be carried by a unique VC associated with that ELAN pair.

We assume that traffic presented by an ELAN for transport on a VC originates as ordinary Ethernet frames, which are variable in size, but no larger than 1518 octets, and also that the average ELAN-to-ELAN traffic on a given VC is 1 MBPS (a fraction of the 10 MBPS total ELAN bandwidth). We also assume an ATM adaptation layer (AAL) structure which segments an Ethernet frame into ATM cells in such a way that the 48 octet cell payload is partitioned into 44 octets to carry user information plus 4 octets to carry AAL control information. The average rate at which information-carrying cells must be transported on a VC is then 2.84 cells/millisec., although they originate in bursts of size up to 35 (for the maximum 1518 octets in an Ethernet frame). Finally, we also assume a flow of AAL-control-containing cells (the purpose of which will not be dealt with here) equal to 10% of the flow of information-carrying cells, or .28 cells per millisec.

Moreover, we assume that the network applies BWM controls to these VCs, using the following leaky-bucket parameter values: $c = 4$ cells/millisec., and a value of M in the range $10 \leq M \leq 50$. Thus, the network agrees to transport an average throughput roughly 28% greater than the VC long-term average throughput of 3.12 cells/millisec., but the allowed burstiness may be severely restricted. Note that the 35 cells from a maximum-sized Ethernet frame will take slightly less than 100 microsec. to enter the network, if they are sent at the full 150 MBPS access rate.

There are several options from which a TA could choose in sending ATM cells; two options that illustrate flow controlling possibilities are:

i. For every Ethernet frame, send all the resulting ATM cells at the full 150 MBPS rate into the network. This will typically result in many of the ATM cells being violation-tagged, but could be a good TA strategy if the network is only lightly loaded, depending on how easily recovery from cell discard can be effected.

ii. For every Ethernet frame, separate successive ATM cells by an idle time of τ millisec. This can be thought of as a "traffic-shaping" form of flow control at the TA. In particular, if the choice $\tau = .250$ millisec. is made, then none of the ATM cells will be violation-tagged, but the transfer of an entire Ethernet frame may be unnecessarily slowed down. The operation of such a "traffic-shaper" can be thought of as a single-server queue with a deterministic service time, τ.

4. TRAFFIC/PERFORMANCE MODELS UNDERLYING BWM

A global model or simulation would capture significant interactions. However, it would be too complicated to capture all interactions, so that sub-models are needed for additional insight. It can be seen that the overall control architecture can be viewed as four separate control components, each of which can be studied through a separate model or set of models. In particular,

i. A SCD (selective cell discard) network node model would focus on the performance characteristics of a simple discard mechanism for violation-tagged cells, e.g., based on an instantaneous local queue threshold, and would allow the cell traffic entering the node to be broken into its two components (violation-tagged and non-violation-tagged), with each component separately characterized by its intensity and a measure of burstiness, e.g., *peakedness*[13]. This model would allow quantification of delays and losses for non-violation-tagged and violation-tagged cells individually. The inputs for a SCD analysis module would be obtained as the outputs from other SCD modules as well as TBF1,2 modules described below.

ii. A pair of models, TBF1 and TBF2 (TBF = throughput-burstiness filter), would characterize the traffic impact of throughput-burstiness filtering through the leaky-bucket traffic monitoring. These models would allow the VC cell stream being filtered to be characterized by its original throughput and burstiness, and would produce the resulting throughputs and burstiness parameters of the two resulting cell streams: violation-tagged and non-violation-tagged. TBF1 focuses on the resulting non-violation-tagged traffic, and TBF2 focuses on the violation-tagged traffic.

iii. In an overall network congestion control analysis procedure, the outputs of TBF1 and TBF2 modules would feed a network of SCD modules, to provide a quantification of the congestion-dependent edge-to-edge transport characteristics of the network.

iv. Finally, given this edge-to-edge transport characteristic, i.e., for the "statistical VC" mentioned in Section 2.3, the end-to-end flow and error controls can be studied. For a specific set of end device controls, the corresponding FEC (flow/error control) module would couple with this edge-to-edge transport characteristic to quantify overall end-to-end performance.

Related papers[14] [15] address methodologies for characterizing the module TBF1, and ongoing activities are being similarly focused at the remaining modules, as well as analytic methods

for coupling these modules together.

REFERENCES

1. "Broadband Aspects of ISDN," *CCITT COM XVIII Draft Recommendation I.121*, June, 1988.
2. "Meeting Report of Sub-working Party 8/1—ATM," *CCITT COM XVIII-TD.14-E*, Geneva, June, 1989.
3. N. Shacham, "Packet Recovery and Error Correction in High-Speed Wide-Area Networks," to be presented at MILCOM'89.
4. D.W. Petr, L.A. DaSilva, Jr., and V.S. Frost, "Priority Discarding of Speech in Integrated Packet Networks," *IEEE J. on Sel. Areas in Comm.*, Vol. 7, No. 5, pp. 644-56, June, 1989.
5. C. Chamzas and D.L. Duttweiller, "Encoding Facsimile Images for Packet-Switched Networks," *IEEE J. Sel. Areas Comm.*, June, 1989.
6. F. Kishino, K. Manabe, Y. Hayashi, and H. Yasuda, "Variable Bit-Rate Coding of Video Signals for ATM Networks," *IEEE J.Sel. Areas Comm.*, June, 1989.
7. R.G.H. Rogers, P.S. Richards, and G.M. Woodruff, "An Assessment of Network Control Configurations for a Packetized Multimedia Communications Network," *IEEE COMSOC Int'l. Workshop on Future Prospects of Burst/Packetized Multimedia Communications*, Osaka, 1987.
8. G.M. Woodruff, R.G.H. Rogers, and P.S. Richards, "A Congestion Control Framework for High-Speed Integrated Packetized Transport," *Proceedings of the IEEE GLOBECOM'88*, November, 1988.
9. G. Gallassi, G. Rigolio, and L. Fratta, "ATM: Bandwidth Assignment and Bandwidth Enforcement Policies," to be presented at IEEE Globecom'89.
10. D. Sparrell, "Wideband Packet Technology," *Proceedings of the IEEE GLOBECOM'88*, November, 1988.
11. J.S. Turner, "New Directions in Communications (or Which Way to the Information Age?)," *IEEE Communications Magazine*, October, 1986.
12. R.L. Cruz, "Maximum Delay in Buffered Multistage Interconnection Networks," *Proceedings of the IEEE INFOCOM'88*, April 1988.
13. A.E. Eckberg, "Generalized Peakedness of Teletraffic Processes," *Proceedings of the 10th International Teletraffic Congress*, Montreal, 1983.
14. A.E. Eckberg, D.T. Luan, and D.M. Lucantoni, "Bandwidth Management: A Congestion Control Strategy for Broadband Packet Networks — Characterizing the Throughput-Burstiness Filter," Int'l. Teletraffic Congress Specialists Seminar, Adelaide, 1989.
15. A.E. Eckberg and D.M. Lucantoni, "A Traffic/Performance Analysis of the Bandwidth Management Throughput-Burstiness Filter," to be submitted for publication.