A B-ISDN/ATM TRAFFIC DESCRIPTOR, AND ITS USE IN TRAFFIC AND CONGESTION CONTROLS

A. W. BERGER and A. E. ECKBERG

AT&T Bell Laboratories
P.O. Box 3030, Holmdel, NJ 07733-3030, USA

Abstract

This paper summarizes issues underlying the need for a *traffic descriptor* methodology for B-ISDN/ATM, and identifies objectives to drive the selection of an appropriate traffic descriptor. Two of these objectives are that the traffic descriptor should relate strongly to: (i) B-ISDN/ATM Call Admission Control (CAC), and (ii) B-ISDN/ATM Usage Parameter Control (UPC). Motivated by these two objectives, a traffic descriptor related to *traffic peakedness* is investigated and shown to have considerable promise.

1. Introduction

This paper reports work in progress on the topic of a B-ISDN/ATM traffic descriptor. The study is not yet finalized, and conclusions on this topic have not yet been drawn. This work is being reported now to illustrate what appears to be a promising approach, and to invite others to build on these ideas.

Broadband ISDN (B-ISDN) networks, utilizing Asynchronous Transfer Mode (ATM), are expected to provide the information transport for a rich mixture of services and applications, associated with which will be a broad spectrum of traffic types and transport performance needs. ATM provides a flexible means for supporting a continuum of transport rates, and also provides a potential efficiency from the statistical sharing of network resources (bandwidth, buffers, processing, etc.) by multiple users. To be competitive with specialized high-speed private network alternatives, B-ISDN/ATM networks will need to be engineered to fully exploit this potential for efficiency. With the B-ISDN/ATM goals of supporting diverse service and traffic mixes, and of efficient network resource engineering, the design of a congestion control becomes an important challenge. It is widely held that the ability to meet this challenge will greatly influence the ability of B-ISDN/ATM to compete, and thus the eventual viability of B-ISDN/ATM.

1.1 The Need for a Traffic Descriptor

The need for quantitatively addressing traffic-related issues pervades the subject of B-ISDN/ATM congestion control. Three key areas where traffic will play a fundamental role are:

i. Specifying the traffic-related terms of *service contracts* between the network and end-terminals; i.e., specifying the characteristics of traffic that the end-terminals can expect to be supported by the network, and, conversely, what region of traffic the network can expect end-terminals to stay within.
ii. Providing a key basis for *preventive* congestion control components, in particular, for traffic-based call acceptance/denial strategies.
iii. Providing a framework for *reactive* congestion control components; in particular, a means for ensuring network resource protection and fair resource allocation through real-time, fast-acting traffic-based controls.

Although there has been much debate and little agreement recently on the characteristics needed in an overall B-ISDN/ATM congestion control architecture, there is one area of almost universal agreement: that an effective congestion control architecture must have a means to be traffic-based. A need that has been recently recognized and stressed (e.g., [1]) is that for a *traffic descriptor*, i.e., a framework within which traffic can be quantitatively described. An example, although possibly not the most useful, is a set of traffic parameters such as peak rate, long-term average rate,

CH2980-1/91/0000-0266 $1.00 © 1991 IEEE

burst identifier (e.g., "bursting" may mean sending at a rate that is at least a specified percentage of the peak rate), average per-burst duration, and average inter-burst time.

Some of the desirable characteristics of a traffic descriptor are (see, e.g., [1]):

- It should be as simple and understandable as possible, ideally not immersed within an abstract mathematical framework.
- It should be useful for specifying traffic-related aspects of service contracts between networks and terminals.
- It should lend itself to a quantitative, and fairly accurate, prediction of the performance impacts, on shared network resources, of a call with given traffic characteristics that the network may or may not admit; i.e., there needs to be an effective link between the traffic descriptor and *Call Admission Control* (CAC).
- It should relate naturally to a simple means for the network to perform "traffic monitoring and policing," also now called *Usage Parameter Control* (UPC), i.e., monitoring the traffic of each admitted call in real-time, and discriminating between *excessive traffic* (outside the terms of the service contract) and *nonexcessive traffic* (within the terms of the service contract). It is important in this regard for both the network and the terminals to be able to test for such *traffic compliance*, where compliant traffic is defined in terms of the traffic descriptor.

That is, an ideal traffic descriptor would tie together, within a common, practical framework, these key ingredients of an overall B-ISDN/ATM traffic management and congestion control.

1.2 Peak Rate as a Traffic Descriptor

While the need for a traffic descriptor has been well recognized, agreement on what that descriptor should be has been difficult to achieve. This is due to the need for the desirable characteristics listed above, which are proving difficult to attain. Thus, agreement has been reached in CCITT (see [1-3]) to focus initially on peak traffic rate as a descriptor. Peak rate is an understandable and unambiguously defined traffic parameter that can be verified in real-time. However, using peak rate as the traffic descriptor may lead to an overly conservative CAC. Thus, while peak rate will be initially used, more encompassing, but still simple, traffic descriptors will be sought.

1.3 Statistical vs. Operational Traffic Descriptors

Viewing a traffic descriptor as a means for specifying user traffic levels that will place demands on shared network resources, there are two fundamentally different approaches that can be taken: (i) a *statistical* approach, and (ii) an *operational* approach:

- The statistical approach is the more conventional in traffic theory, and focuses on statistical traffic parameters, such as the long-term average rate and burst parameters mentioned earlier. Such an approach may be motivated by the apparent availability of methodologies for predicting performance of a network stressed by traffic with given statistical parameters, i.e., for addressing the CAC issue. However, associated with a statistical framework for traffic is the difficult task of verifying a set of statistical traffic parameters. By the very nature of their definitions, statistical traffic parameters may require a lengthy observation interval for verification, making real-time traffic-compliance-testing nearly impossible. See [4] for examples of such difficulties.
- With the operational approach, little heed is given to traffic at levels well within compliance with traffic contracts; rather focus is placed on traffic that is either just compliant or not compliant. This is achieved by *equating an operational traffic descriptor with a parameterized compliance-testing algorithm* intended to discriminate "excessive" traffic from "non-excessive" traffic. Thus, by focusing on traffic-compliance-testing itself, we immediately resolve the issue of real-time traffic-compliance-testing at the terminal; and algorithms for traffic-compliance-testing by the network, i.e., UPC, are immediately suggested. Worst-case traffic patterns that will be deemed compliant (and thus that the network must handle), can still be predicted, once a UPC algorithm has been selected by the network, forming the basis for network CAC algorithms.

1.4 Summary of the Paper

This paper addresses the need for such a traffic descriptor by demonstrating that *traffic peakedness*, a traffic characteristic that has seen much use in traditional traffic engineering (e.g., in trunking problems), captures many of the traffic-related issues that are important in B-ISDN/ATM congestion/traffic controls. Key contributions of the paper are in identifying a methodology that can lead to:

- simple peakedness-based performance-predicting approximations relating the traffic descriptor to CAC strategies,
- a real-time traffic-compliance-testing algorithm that can form the basis of an operational traffic descriptor combining traffic intensity and peakedness in a natural way.

2. An Approach to Treating the Traffic Descriptor, UPC, and CAC

Because the traffic descriptor must relate logically to both the UPC and CAC control capabilities, we now review some aspects of congestion control. Specifically, it is currently envisioned ([3]) that an option of the UPC operation will be as shown in Figure 1.

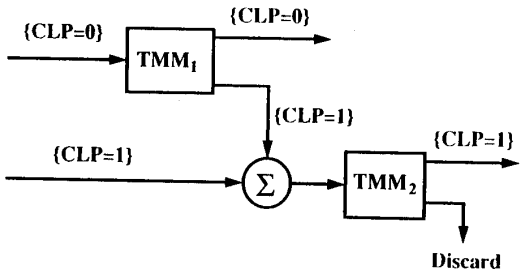

Figure 1. Operation of the UPC

In this figure, "CLP" denotes "cell loss priority," and traffic originating at a terminal is shown entering at the left at both high priority (CLP=0) and low priority (CLP=1). The high priority traffic is monitored with a traffic-monitoring module, TMM_1, for compliance with the traffic contract, and compliant traffic remains at the CLP=0 setting. Non-compliant CLP=0 traffic is set to CLP=1, merged with the original CLP=1 stream, and monitored with module TMM_2, which detects and discards excessive CLP=1 traffic. The key point to be observed is that it is the CLP=0 traffic emerging at the right side of Figure 1 that the network needs to be able to handle. That is, admission of a call, i.e., CAC, needs to be based on the traffic descriptor of this resulting CLP=0 traffic. This is illustrated in Figure 2 (where "TD" denotes "traffic descriptor").

Similarly, it is clear that the UPC needs to relate in a logical way to the traffic descriptor at the source. Note that, even if the source is described by an operational traffic descriptor, i.e., a traffic-compliance-testing

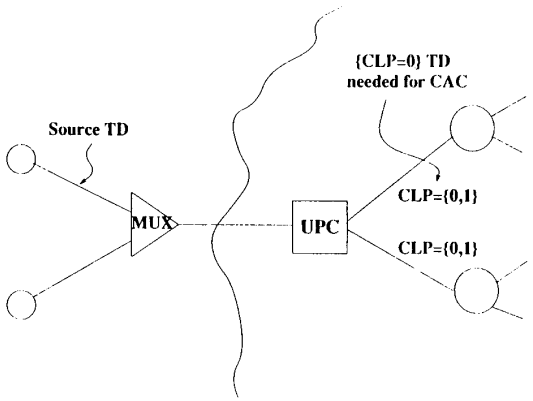

Figure 2. Relation between the source traffic descriptor, the UPC, and the CAC

algorithm, this algorithm may not be appropriate for use as the UPC algorithm. For example, access jitter introduced by possible access multiplexing contention may preclude the UPC using the same algorithm. However, the UPC algorithm needs to strongly relate to the source operational traffic descriptor.

3. The Role of Traffic Peakedness

Traffic peakedness is a traffic characteristic that has proven useful in several performance-predicting applications over the years. It was originally motivated as a measure of the burstiness of call attempt overflows from a primary trunk group, and led to approximate methodologies for predicting call blocking in a circuit-switched network. However, peakedness has been shown ([5]) to be equivalent to any second-order traffic characterization (i.e., those characterizations that capture the correlational structure of a traffic stream), and also to be useful in quite simple delay approximations ([6]).

A useful definition of traffic peakedness for our purposes is as follows. For a traffic stream with intensity λ and *expectation function* $U(\cdot)$, defined as:

$U(x)$ = *the expected number of arrivals following, and no later than a time x from, an arbitrarily chosen arrival,* for $x \geq 0$
= 0, for $x < 0$

the peakedness with respect to parameter β, $z(\beta)$, is given in terms of λ and $\hat{U}(\cdot)$, the Laplace-Stieltjes transform of $U(\cdot)$,

$$\hat{U}(s) = \int_{0-}^{\infty} e^{-sx} dU(x)$$

as

$$z(\beta) = 1 + \hat{U}(\beta) - \lambda/\beta$$

If traffic with intensity λ and peakedness $z(\cdot)$ is handled by a single-server queue with first and second service time moments τ_1 and τ_2, then the resulting complementary delay distribution can be approximated as

$$P[W>x] \approx \alpha\, e^{-\beta x}, \text{ for } x \geq 0$$

where α and β are determined as

$$\beta = \frac{1-\rho}{\tau_r + \tau_1(z(\beta) - 1)}$$

$$\alpha = 1 - \frac{\beta\tau_r}{\beta\tau_r z(\beta\tau_r/\tau_1) + \rho}$$

and where the parameter $\tau_r = \tau_2/(2\tau_1)$, and $\rho = \lambda\tau_1$. This approximation is derived in [6]; enhanced peakedness-based approximations are currently being worked.

From the above, it can be seen that the parameter β, originally introduced as an independent parameter in the definition of peakedness, has more significance: taking β equal to the decay rate in a queueing system results in the peakedness value, $z(\beta)$, that best captures the burstiness of the traffic that is interacting with that queueing system.

From the type of approximation given above, one can approximately characterize the set of (λ,z) combinations that would result in acceptable delays at a system; it would be represented as in Figure 3, wherein acceptable system utilization (ρ_0) and traffic peakedness (z_0) combinations due to CLP=0 traffic lie under the curve. Figure 3 represents the qualitative characteristic of a queueing component in a network. Clearly, more detail is required for quantitatively characterizing a particular queueing component. Moreover, focus can also be brought to ATM cell loss, as well as delay, characteristics by using peakedness-based performance approximations.

We now focus attention at the UPC, and note that of interest is the traffic characterization of the exiting CLP=0 traffic (as noted in Figures 1 and 2). In particular, to determine whether the CAC should admit or deny a call request, focus needs to be brought to the most extreme CLP=0 traffic that can exit the UPC. The traffic characteristics of this exiting CLP=0 traffic

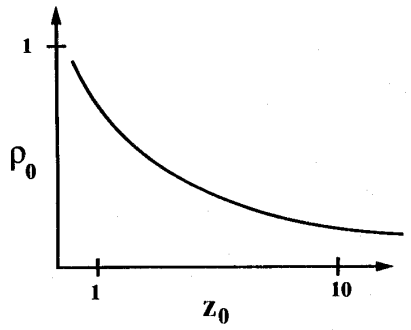

Figure 3. Acceptable combinations of peakedness and utilization

would appear qualitatively as in Figure 4, where $\lambda_{filt,0}$ and $z_{filt,0}$ denote the traffic intensity and peakedness.

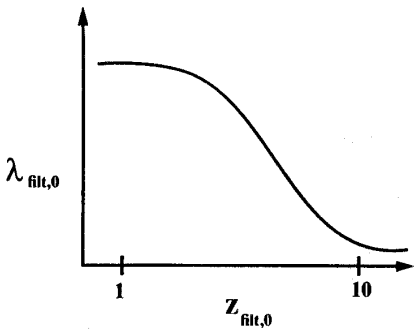

Figure 4. Intensity and peakedness of CLP=0 traffic exiting a UPC

Roughly speaking, a call with associated UPC and extreme traffic profile as given in Figure 4 should be admitted by the CAC only if the total traffic from this call and all other already-setup calls will result in (ρ_0,z_0) combinations under the curve in Figure 3. There are simple means for combining the $(\lambda_{filt,0},z_{filt,0})$ characteristics of all accepted calls to predict their combined impact on the network resources. Also, note that the UPC traffic profile, as depicted in Figure 4, is only one possible profile. For different UPC algorithms, the traffic profile may fall more rapidly, and "fit" better into the acceptable region of Figure 3.

4. Real-Time Peakedness Measurement

A recent observation with respect to peakedness that has potential importance to the subject of traffic descriptors is as follows. Consider observing a traffic stream in real-time, and devising a means for real-time estimation of the traffic intensity and peakedness. Let $\{t_n\}$ denote the sequence of inter-arrival times, and consider observing N such arrivals following an arrival at time $t=0$. Clearly, an estimate of the traffic intensity is just

$$\lambda \approx \left[\frac{1}{N} \sum_{i=1}^{N} t_i \right]^{-1}$$

Moreover, it can be shown that an estimate of the quantity $\hat{U}(\beta)$, from the observed $\{t_n\}$ is just

$$\hat{U}(\beta) \approx \frac{1}{N} \sum_{i=1}^{N} x_i$$

where the sequence $\{x_n\}$ is defined recursively as

$$x_0 = 0$$

$$x_n = e^{-\beta t_n} (1 + x_{n-1}), \text{ for } n>0$$

The above can be easily shown by considering the Laplace-Stieltjes transform of the sample counting function associated with the arrivals, which relates to the expectation function $U(\cdot)$, and can be seen to be asymptotically unbiased as N becomes large. This result, together with the definition we have given in this paper for peakedness, then provides a real-time method for estimating peakedness from actual inter-arrival times.

5. A Peakedness-Based Operational Traffic Descriptor

Upon examining the above, it can be seen that the sequence $\{x_n\}$, computed in real-time at successive arrival epochs, contains information that combines both the traffic intensity and the traffic peakedness. That is, the sample average of the quantities x_n is, from the foregoing, just asymptotically equal to $\hat{U}(\beta) = z(\beta) - 1 + \lambda/\beta$, i.e., a linear combination of the intensity and peakedness. This suggests using as a potential traffic-compliance-testing algorithm the following:

i. initialize x_0 to 0
ii. at the nth arrival, following interarrival time t_n, compute $a_n = e^{-\beta t_n}$
iii. compute $x_n = a_n (1 + x_{n-1})$
iv. if $x_n < M$, declare the nth arrival to be traffic-compliant
v. otherwise, recompute $x_n = a_n x_{n-1}$, and declare the nth arrival to be non-traffic-compliant

It can be seen that this operational traffic descriptor has some desirable characteristics, e.g., it allows the highest intensity with smooth traffic (i.e., constant spacing between arrivals), and requires reduced intensity as traffic burstiness increases. Also, it depends on a single negotiation parameter, M, which combines negotiated intensity and burstiness. The parameter β would typically be chosen as the delay decay rate in a network component at engineered load; a typical value of β^{-1} might be 100 ATM cell emission times. For example, with this value of β, a negotiated value of M equal to 9.508 would allow the following periodic traffic patterns and average rates:

periodic cell pattern	rate
1 cell every 10 slots	.1000
2 cells every 21 slots	.0952
5 cells every 63 slots	.0794
8 cells every 133 slots	.0602

REFERENCES

[1] CCITT Special Rapporteur on Networking and Resource Management, "Report of the Meeting of SWP 8-7," Temporary Document 43 (XVIII/8), Matsuyama, Nov.-Dec., 1990.

[2] CCITT Special Rapporteur SWP XVIII/8-7, "Report of the Meeting," Temporary Document 36 (XVIII/8), Geneva, June, 1991.

[3] CCITT Special Rapporteur SWP XVIII/8-7, "Traffic Control and Resource management in B-ISDN," Draft Recommendation I.trf, Paris, October, 1991.

[4] A.W. Berger, A.E. Eckberg, T.-C. Hou, and D.M. Lucantoni, "Performance Characterizations of Traffic Monitoring, And Associated Control, Mechanisms for Broadband 'Packet' Networks," *Proc. GLOBECOM'90*, 1990.

[5] A.E. Eckberg, "Generalized Peakedness of Teletraffic Processes," *Proc. 10'th International Teletraffic Congress*, 1983.

[6] A.E. Eckberg, "Approximations for Bursty (and Smoothed) Arrival Queueing Delays Based on Generalized Peakedness," *Proc. 11'th International Teletraffic Congress*, 1985.