Performance Characterizations of Traffic Monitoring, and Associated Control, Mechanisms for Broadband "Packet" Networks

A.W. Berger
AT&T Bell Laboratories
Room 3H-601
Holmdel, NJ 07733 USA

A.E. Eckberg
AT&T Bell Laboratories
Room 3J-611
Holmdel, NJ 07733 USA

Ting-Chao Hou
AT&T Bell Laboratories
Room 3K-602
Holmdel, NJ 07733 USA

D.M. Lucantoni
AT&T Bell Laboratories
Room 3K-601
Holmdel, NJ 07733 USA

Abstract

This paper addresses a number of performance issues associated with real-time traffic-monitoring schemes, as will likely be used in broadband "packet-transport-based" networks such as Broadband ISDN (B-ISDN) with Asynchronous Transfer Mode (ATM). The issues of: an appropriate performance framework for traffic-monitoring scheme comparisons; appropriate control actions based on traffic-monitoring; and how traffic parameters (that could enter strongly into "service contracts" between the network and users) should best be specified are treated in an overview summary. Some of the more commonly suggested traffic-monitoring schemes are described, and the leaky-bucket scheme is analyzed. Also, a relationship between traffic *peakedness* and traffic-monitoring schemes is introduced.

1. Introduction

The connection-oriented information transport mechanism to be provided by B-ISDN/ATM networks will allow great flexibility in accommodating a wide spectrum of services and applications, all with different traffic characteristics. It has been recognized that an overall congestion, flow, and error control architecture will be needed for such networks, together with the end applications that will be supported by these networks. And, it has been generally recognized that within this overall control architecture an essential element will be the ability for the network to monitor traffic on connections in real time.[1] The purpose of such traffic-monitoring is to compare real-time estimates of traffic characteristics on a connection with traffic characteristics that have been agreed to between the network and the end-devices of the connection, at the time of the connection set-up, and, depending on the outcome of such a comparison, for the network to take appropriate control actions.

This paper address, in a summary manner, some of the issues associated with real-time traffic-monitoring. More detailed results, in the areas briefly discussed in this paper, can be obtained upon request from the authors.

2. Control Actions Based on Traffic-Monitoring

It is important to observe at the outset that some of the performance objectives associated with real-time traffic-monitoring schemes will depend on what control actions might be stimulated by the result of the traffic-monitoring. There are two general classes of control actions possible when "excessive" traffic has been detected:

i. *policing,* where the network unconditionally discards the excessive traffic after its detection,
ii. *traffic-violation-tagging,* where the excessive traffic is identified and marked, to be discarded later in the network only if congestion is encountered.

Policing tends to be a "hard" control action, while violation-tagging is "softer," in that, depending on the network congestion conditions, very little violation-tagged traffic may actually be discarded. Because the eventual fate of "excessive" traffic depends on which of these control actions is to be taken, one can expect that different objectives would be placed on the accuracy of the traffic-monitoring, depending on whether a policing or a violation-tagging approach is taken. In particular, because violation-tagging is "softer," one can afford to incorrectly identify "non-excessive" traffic as

CH2827-4/90/0000-0350 $1.00 © 1990 IEEE

"excessive" more frequently.

3. A Performance Framework

Of equal importance to the issue of triggered control actions is the question of determining the essential performance parameters of a traffic-monitoring scheme; i.e., determining the framework within which one scheme might be compared with another. We focus on three "dimensions" capturing the key elements of the performance of a monitoring scheme:

i. the scheme's *responsiveness,* meaning how rapidly changes in traffic characteristics into the "excessive traffic domain" can be detected;

ii. the scheme's *probability of false alarm,* meaning with what probability segments of stationary, and non-excessive, traffic will be incorrectly identified as "excessive;"

iii. the amount of *margin* that the scheme requires (Because we would like a monitoring scheme to be responsive, but not exhibit a high probability of false alarm, there is a "margin" that a scheme must allow in traffic above and beyond the agreed-to traffic characteristics; this "margin" can potentially be exploited by a sophisticated and malicious end-device so as to send traffic consistently at an excessive rate, but such that none of the traffic is identified as excessive).

One would want a monitoring scheme to be responsive, to have a low probability of false alarm, and to need only a modest "margin." Clearly, these three performance parameters are involved in basic performance tradeoffs, and it would be impossible to simultaneously achieve very small response time, very small probability of false alarm, and a very small margin. For example, to achieve a very small probability of false alarm with a very small margin would require considerable traffic averaging over long intervals, and this would result in poor responsiveness.

These three performance "dimensions" are involved in key tradeoffs in the overall performance of a traffic-monitoring scheme, and quantifying these tradeoffs for alternative monitoring schemes would form the basis of performance comparisons.

4. Previous Studies on Traffic-Monitoring Schemes

Numerous monitoring schemes have been proposed and analyzed in the literature.[2] [3] [4] Four of them, the leaky-bucket scheme, the jumping window scheme, the moving window scheme, and the exponential smoothing scheme, are briefly described below.

In the leaky-bucket scheme, there is a counter associated with each virtual connection. The counter is incremented upon cell arrivals and decremented at a constant rate (c) as long as the counter value is positive. It is flexible in the sense that either the average rate or the peak rate can be monitored. The counter value reflects deviation from the target measure. If counter exceeds a predetermined threshold (i.e., bucket size), the cell is either discarded or tagged as droppable. For average rate monitoring, a higher threshold is required for bursty traffic. The implementation of the leaky-bucket scheme is simple. It requires one counter for the bucket content and two variables for the decrement interval and the counter limit.

In the jumping window scheme, the counter starts at zero, at the beginning of a fixed time interval (window). The counter is incremented upon cell arrivals within the window, and resets to zero at the end of the window. The next window starts immediately after the preceding window. If the counter exceeds a predetermined threshold, the cell is either discarded or tagged as droppable. For average rate monitoring, a larger window size is required for bursty traffic. The implementation complexity of this mechanism is comparable to that of the leaky-bucket scheme.

In the moving window scheme, the maximum number of cell arrivals within a window of fixed time interval (W time slots) is limited as in the jumping window scheme. The difference is that adjacent windows overlap by $W-1$ slots. Thus, each cell has effect on a total number of windows equal to W. It therefore has higher implementation complexity since it requires that the arrival time of a cell be stored for W slots.

In the exponential smoothing scheme, as in the jumping window scheme, consecutive windows do not overlap. However, the threshold in the i-th window is a function of an exponentially weighted sum of the number of accepted cells in the preceding windows. If no weight is given to previous windows, it degenerates to the jumping window scheme. The time complexity of this scheme is higher than the above three schemes, but it does not require as much storage space as the moving window scheme.

5. Performance of the Leaky-Bucket Monitor

In this section we illustrate how a traffic monitoring scheme can be characterized in the framework described in Section 3. We use the leaky-bucket scheme as an example. Figure 1 is the 3-dimensional characterization of the leaky bucket scheme when used to monitor a source characterized by an on-off model. The on-off source model shown in Figure 2 has been used in many performance studies* for a single packetized voice source,[5] [6] or for a bursty data source.[7] It is characterized by the mean number of cells per talkspurt, the minimum inter-cell spacing, and the mean silence duration. Assume that the mean talkspurt duration is 354 msec, and the mean silence duration is 650 msec. Since every 6 msec (48 bytes) of voice forms a cell, the mean number of cells per talkspurt (or burst) is 59. It is also assumed that the minimum inter-cell spacing is 6 cell time. This corresponds to a scenario where the voice cells are sent over a transmission link running at 6×53/48=6.625 times of the source peak rate. Given a margin M, and a bucket size B, we can obtain the cell tagging probability by solving a discrete $G/D/1/B-1$ queue.[2]

Note that in Figure 1, one of the dimensions is the non-responsiveness (NR), instead of the responsiveness listed in Section 3. Choosing non-responsiveness has the advantage of conforming better to the other two dimensions in that being closer to the origin is the better. Here the non-responsiveness is defined as the longest burst of cells that can pass through the leaky-bucket without being tagged.

The surface in Figure 1 defines the feasible performance region of the leaky-bucket scheme. We observe that the probability of false alarm (PFA) is significant (say, greater than 10^{-3} for voice) if either M or NR is small. The PFA is less than 10^{-3} when both $M>1.0$ and $NR>50$. The exception to the above is when the margin M is greater than a critical value (approximately 1.5 in this example). In that case, the PFA is always zero due to the fact that the decrement interval (inverse of the drain rate) of the leaky-bucket is equal to or less than the minimum inter-cell spacing.

The above example has shown that it is difficult for the leaky-bucket scheme to accurately monitor a packetized voice source. For burstier sources, like high speed file transfers, the mean burst is larger, and the minimum inter-cell spacing is smaller. It will require much larger margin and non-responsiveness to achieve the more stringent (say, 10^{-9}) cell loss requirement. Other studies, e.g, [2] , also showed that most monitoring schemes can not accurately distinguish between excessive and non-excessive traffic. Although a more sophisticated monitoring scheme may be able to do a better job, it generally gets too complicated to be implemented at B-ISDN speeds. If we accept that incorrectly identifying "non-excessive" traffic as "excessive" is inevitable, then a soft control action based on the results of the traffic-monitoring definitely is better than a hard control action. That is, in the terminology of Section 2, traffic-violation-tagging is preferable to policing.

6. A Fourth Traffic-Monitoring Scheme

Another traffic-monitoring scheme is motivated by some classical results in characterizing bursty traffic. Specifically, it has been shown that the *peakedness* characterization captures in a fairly robust way the "burstiness" and/or the "smoothness" of a traffic stream;[8] moreover, the peakedness characterization lends itself to some simple performance approximations for quantities such as delays and losses from finite-capacity queues,[9] performance parameters of interest when addressing broadband transport. It is shown in [8] that the peakedness of a stream is equivalent to the long-term average, taken over all cell arrival epochs, of the output of an exponentially-weighted summing system. See Figure 3. This summing system, itself, then suggests a possible traffic-monitoring scheme that is directly related to monitoring traffic peakedness.

This possible approach to traffic-monitoring will be expanded upon in follow-up papers.

7. Some Additional Observations

It has been suggested in Section 2, and quantified in Section 5, that traffic-monitoring cannot be performed with a high level of performance in all dimensions, sufficient to allow policing (i.e., unconditional discard of traffic identified to be "excessive") to be done. Thus, a softer action, such

* The parameters of this example are for relatively low-speed traffic connections. However, these parameters can be easily scaled to give insight into higher speed connections that may be more common in a broadband environment.

as traffic-violation-tagging is preferred. Another "soft" control action is to exploit that fact that when monitoring a set of traffic streams, some streams may be momentarily considered to be "excessive" while others may be at low activity. The latter streams may be considered to have accumulated "credits" in their traffic-monitors which could be "shared" by the former streams to allow avoiding traffic-monitoring-triggered control actions. Such ideas are developed more fully in [10].

More fundamentally, by quantifying the tradeoffs between the performance "dimensions," we have brought focus to ambiguities inherent in the use of a finite set of traffic characteristics (e.g., average rate, peak rate, "burstiness") to accurately distinguish between excessive and non-excessive traffic. We are then led to the conclusion that to minimize such ambiguity the best approach would be for the network and individual end-devices (whose traffic is to be monitored) to use the traffic-monitoring algorithm (and its specific set of parameters) itself as the basis for defining allowable (i.e., non-excessive) traffic characteristics.

REFERENCES

1. A. E. Eckberg, D. T. Luan, and D. M. Lucantoni, "An approach to controlling congestion in ATM networks," *Int. J. of Digital and Analog Communication Systems,* Vol. 3, 199-209, 1990.
2. E.P. Rathgeb, "Policing Mechanisms for ATM Networks - Modeling and Performance Comparison," *Proc. 7th ITC Seminar on Broadband Technologies: Architectures, Applications, Control and Performance,* October, 1990.
3. P. Joos and W. Verbiest, "A Statistical Bandwidth Allocation and Usage Monitoring Algorithm for ATM Networks," *Proc. IEEE ICC,* 1989.
4. L. Zhang, "A New Architecture for Packet Switching Network Protocols," Ph.D. Dissertation, MIT, 1989.
5. K. Sriram and W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data," IEEE JSAC, Vol. SAC-4, N0. 6, September 1986.
6. H. Heffes and D. M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," IEEE JSAC, Vol. SAC-4, N0. 6, September 1986.
7. T.-C. Hou and A. K. Wong, "Queueing Analysis for ATM Switching of Mixed Continuous-Bit-Rate and Bursty Traffic," Proc. IEEE INFOCOM 1990.
8. A. E. Eckberg, "Generalized Peakedness of Teletraffic Processes," *Proc. 10'th International Teletraffic Congress,* 1983.
9. A. E. Eckberg, "Approximations for Bursty (and Smoothed) Arrival Queueing Delays Based on Generalized Peakedness," *Proc. 11'th International Teletraffic Congress,* 1985.
10. A. Berger and W. Whitt, "A Mulit-Class Input-Regulation Throttle," *Proc. 29th IEEE Conf. on Decision and Control,* December, 1990.

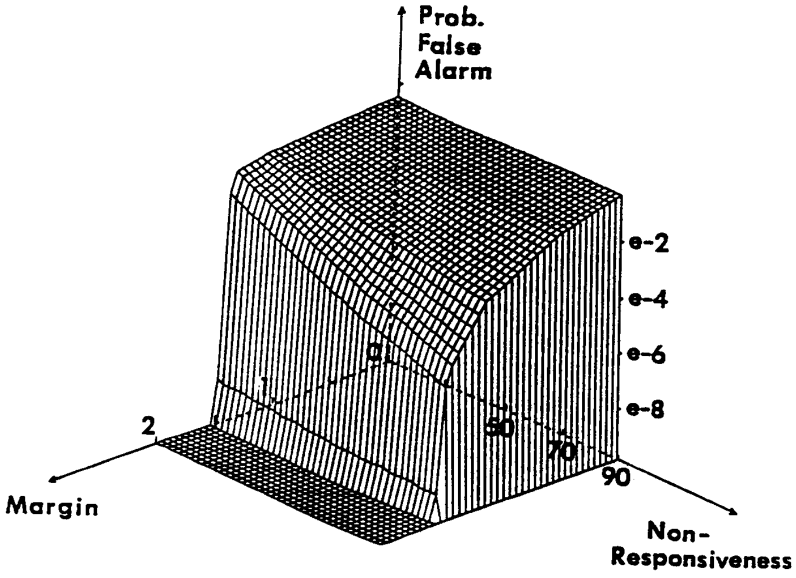

Figure 1. 3-Dimentional Characterizauon of the Leaky-Buckat Scheme.

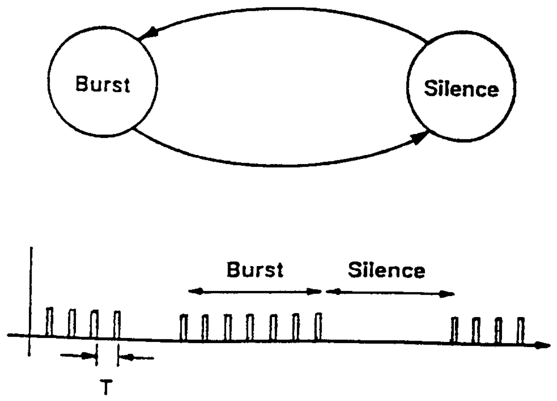

Figure 2. Two-State On-Off Source Model.

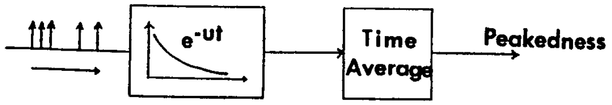

Figure 3. Exponentially-Weighted Summing System and Peakedness.