40
An Efficiency Prediction Method for ATM Multiplexers

Kalyan S. Perumalla[†], *C. Anthony Cooper*[*] *and Richard M. Fujimoto*[†]
[*]*Bell Communications Research, Red Bank, NJ 07701-5699.*
Phone: +1 908 758 2144.
[†]*College of Computing, Georgia Institute of Technology*
Atlanta, GA 30332-0280. Phone: +1 404 894 5620.
Fax: +1 404 894 9442. email: kalyan@cc.gatech.edu

Abstract

This paper describes a methodology for conservatively predicting the traffic carrying efficiency of an ATM multiplexer that is operating in compliance with realistic performance objectives. Our approach uses a multi-phased treatment of this important problem. In the first phase, a synthetic model of bursty traffic sources is used to construct a set of loading curves that predict worst-case statistical multiplexing efficiencies as a function of certain parameters which characterize both the traffic sources and the multiplexing system under consideration. In the second phase, a set of real traffic sources is matched with these same parameters through appropriate traffic measurements, and a prediction of achievable statistical multiplexing efficiency is obtained by the appropriate use of these loading curves. This technique has been successfully tested on real traffic source data obtained from Local Area Networks, and which has been shown to possess the "self-similar" temporal characteristic that is known to present significant challenges for statistical multiplexing. This methodology was developed with the aid of a high-performance, parallel simulator that is used both for the construction of representative loading curves and for demonstrating the conservative nature of the predicted multiplexing efficiencies that result when this methodology is applied to samples of real traffic. When augmented with a suitable set of routine traffic measurements, it is anticipated that the methods described here can play a significant role in practical ATM network dimensioning processes.

Keywords

ATM multiplexers, multiplexing efficiency, parallel simulation, Markov Chain models, Ethernet LAN traces, self-similar

1 INTRODUCTION

The deployment of telecommunications network equipment based upon Asynchronous Transfer Mode (ATM) technology poses a significant challenge with respect to the net-

L. Mason et al. (eds.), *Broadband Communications*

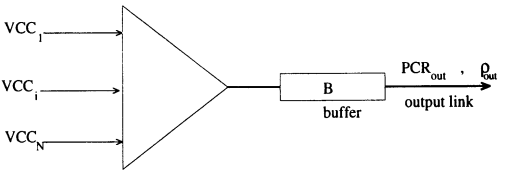
work dimensioning methodology needed to achieve economic levels of network equipment usage while simultaneously maintaining satisfactory levels of network performance. The methodology described in this paper addresses a key element of this ATM network dimensioning challenge — the prediction of ATM multiplexing efficiencies achievable under realistic performance objectives while operating with traffic loads imposed by real traffic sources. This methodology is most directly applicable for dimensioning ATM networks and network equipment used in support of Permanent Virtual Connection applications. When used in combination with additional methods for treating the blocking experienced by ATM connections of diverse capacities, one may reasonably anticipate the extension of this methodology for dimensioning ATM networks supporting Switched Virtual Connection applications.

A variety of approaches have been advanced for addressing this ATM network dimensioning challenge, of which Bensaou (1990), Hui (1990) and Elwalid (1995) could be viewed as representative. In addition, significant evidence exists concerning the adverse performance impacts of improperly dimensioned statistical multiplexers operating on traffic that possesses the self-similar temporal characteristic (Leland (1994), Erramilli (1994)). It is therefore prudent for any ATM dimensioning methodology to include verification of its multiplexer performance predictions by testing with real traffic loads that possess this self-similar property. Such verification is provided for the method presented here. We remark that an extension of this measurement-oriented verification process can establish a basis for using routine traffic measurements in support of ATM network dimensioning.

Our study develops in several phases a method that predicts maximum safe load levels for ATM multiplexers. This methodology is developed using several distinct types of traffic source models and a high capacity simulator that is capable of completing simulation runs on the order of 10^{10} cell transfers in less than one hour of real time (Nikolaidis (1993), Nikolaidis (1994), Fujimoto (1995)). The unique benefits obtainable with this high capacity simulator become apparent when examining the clumped nature of cell losses.

In the first phase of this study, a traffic source model based upon a low order Markov Chain is used in a simulation-based process to develop a set of loading curves that are intended to provide conservative, or worst-case, predictions of multiplexer efficiency. These loading curves are parameterized by certain dimensionless ratios relating to both multiplexer and traffic source characteristics. In the second study phase, predictions of multiplexing efficiency based on these loading curves and suitable parameter matching techniques are validated using results from simulations driven by samples of Local Area Network (LAN) traffic.

The remainder of this paper is organized as follows. First, our goals and approach are described and the rather general statistical multiplexing model used for this study is identified. The Markov Chain-based synthetic traffic source model used in the initial phase of this work is then cited, together with the preliminary simulation-based results used to select "worst case" parameters for this traffic source model. Next considered are simulation techniques based on this initial traffic source model and the generation of loading curves. We then describe the simulation technique used in the second phase of this study, which uses a traffic source model based upon Ethernet LAN traffic samples. Comparison between the multiplexing efficiencies predicted from the loading curves and the multiplexing efficiencies achievable with the LAN traffic samples demonstrates the conservative character of this methodology.



$PCR(VCC_i)$ = Peak Cell Rate of VCC_i, $i = 1, 2, \cdots, N$
PCR_{out} = Peak Cell Rate supportable on output link
$G = \frac{\sum_{i=1}^{N} PCR(VCC_i)}{PCR_{out}}$

Figure 1 Statistical Multiplexer Model used in this Study.

2 GOALS AND APPROACH

This methodology's primary goal is to conservatively predict the statistical multiplexing gain and output link utilization of an ATM statistical multiplexer that can be achieved with real traffic while complying with established performance objectives, particularly for the recognized Cell Loss Ratio (CLR) performance parameter (ANSI Standard (1994)). The traffic of interest here has an intensity that varies randomly in time between zero and a maximum value that is described by a Peak Cell Rate (PCR). Such traffic can include the output of various data or video traffic sources; it has been referred to as Statistical Bit Rate traffic; and certain additional quantifying parameters* for this traffic have been introduced in international standards (ITU (1995)). The CLR performance objective of 10^{-7} is used here as a realistic and previously documented value for the cell loss impairment experienced by a data-oriented application when carried by an ATM virtual connection across one ATM multiplexer (Bellcore (1995)).

A secondary goal of this research is to better characterize the cell loss events occurring at an ATM statistical multiplexer. Such characterization should have sufficient accuracy to support the satisfactory interpretation of simulation results with respect to the observed cell losses on a given simulation run, and the resulting inference about the likely range of CLR values.

Our approach is based upon the use of two distinct types of traffic source models. Each type of traffic source model is used in a separate study phase to represent inputs to an ATM statistical multiplexer. The rather general multiplexer model shown in Figure 1 is used throughout this study. This statistical multiplexer can be viewed as having N inputs, which are interpreted as ATM connections, or more specifically, Virtual Channel Connections (VCCs), and a single output of known capacity that is characterized by its Peak Cell Rate, PCR_{out}, and that is assumed to be supplemented with an output buffer of known size, identified as B cells. The details of internal data transfer between the inputs and the output of this multiplexer are submerged for purposes of this study, and so it is only necessary to assume that any queuing at input buffers is either negligible or else

*Relevant parameters are the Sustainable Cell Rate (SCR), in cells/second, and the Maximum Burst Size (MBS), in cells.

approximated by an equivalent amount of output buffering. It is generally desirable to achieve large values for N and for two related measures of multiplexer efficiency, which are the statistical multiplexing gain and the output link utilization. It is taken as mandatory that the CLR for this system remain less than its objective value, so that all such efficiency measures are maximized subject to this constraint.

Three dependent variables are useful for characterizing the efficiency of an ATM multiplexer. If we let N be the number of inputs, or traffic sources, driving the multiplexer, these dependent variables are:

$$\begin{aligned} N^* &= \text{maximum value of } N \text{ that meets the CLR performance objective for a system configuration} \\ G &= \text{multiplexing gain, which is relatable to } N^* \text{ as } G = \frac{N^* \cdot PCR(VCC)}{PCR_{out}} = \frac{N^*}{K} \\ \rho_{out} &= \text{output link utilization, which is relatable to } G \text{ as } \rho_{out} = \frac{G \cdot SCR(VCC)}{PCR(VCC)} = \frac{G}{b} \end{aligned} \quad (1)$$

where it is convenient to assume that $PCR(VCC_i) = PCR(VCC)$ and $SCR(VCC_i) = SCR(VCC)$ for $i = 1, \ldots, N$, and where K and b are dimensionless ratios whose definitions follow from the relations provided in (1).

The first type of traffic source model used here is based upon a low order Markov Chain that includes sufficient detail to permit independent adjustment of the first order and second order burst length statistics for traffic sources having ON/OFF emission characteristics. A four-state Markov Chain is selected for this purpose. Most of the parameters associated with this source model and the associated multiplexer are next set to "worst-case" values, which are intended to yield the most pessimistic simulation-based results possible for the largest achievable N^*, G and ρ_{out} when the multiplexing system is operating in compliance with its CLR performance objective. Then the remaining parameters of this worst-case configuration (most of which are expressed as dimensionless ratios) are varied to produce a set of loading curves. It is hypothesized that, when the statistics of real traffic sources are appropriately matched with the parameters that index these loading curves, safe predictions can be made concerning the largest values of N^*, G and ρ_{out} that will be achievable.

The second type of traffic source model used in this study incorporates time-based records, or traces, of Ethernet LAN traffic (Leland (1994)). We select this type of traffic for testing the loading curve predictions because such traffic has been shown to possess the self-similar temporal characteristic that makes its statistical multiplexing difficult (Leland (1994), Erramilli (1994)).

Simulation of high speed ATM multiplexers typically requires long execution times, owing to the large number of cell transfer events to be simulated in support of a practical CLR performance objective. While conventional sequential simulators are capable of simulating on the order of 10^7 cell transfers in a reasonable amount of processing time, our results indicate that significantly higher simulation capacity is desired to study cell losses commensurate with a CLR objective of 10^{-7} and to provide adequate consideration for the demonstrably nonindependent nature of cell losses.

Our research makes use of appropriately structured parallel simulation techniques that extend the methods of Fujimoto (1995) and allow about 10^{10} cell transfers to be simulated in about one hour of wall clock time. Our implementation is based on a 32 processor Kendall Square Research shared-memory multiprocessor. This simulator allows the ATM

multiplexer's input links to be driven by a variety of source models. The two source models used in the current study are a four-state Markov Chain model, and a LAN trace-driven model.

3 ROLE OF A FOUR-STATE MARKOV CHAIN SOURCE MODEL

A four-state Markov Chain model having an ON/OFF operating mode is selected as our initial traffic source model because it provides several desirable features. First, it permits the independent adjustment of both first order and second order burst-oriented statistics by suitable treatment of the sojourn times in each of its four states. Its characteristics and application to this type of study are reasonably well understood (Eliazov (1990)). Furthermore, it may support analytically tractable analysis of the queuing aspects associated with the multiplexing system under consideration. Since our goal is to relate intermediate results in the form of loading curves obtained with this source model to the statistical multiplexing efficiencies achieved with real traffic sources (and for which analytically tractable models are not generally available), the focus of this work is oriented towards simulation-based analysis.

In general, each of the dependent variables N^*, G and ρ_{out} will be a function of a number of independent variables that characterize various aspects of the traffic sources and multiplexing system under consideration. The more significant of these independent variables are:

- B = size of output buffer, in cells
- K = capacity factor for a traffic source and this output link = $\frac{PCR_{out}}{PCR(VCC)}$
- CLR_{obj} = Cell Loss Ratio objective = 10^{-7} for this study
- MBS = Maximum Burst Size = bound on a traffic source's active period or burst length
- $c^2(A)$ = squared coefficient of variance of the active (ON) period of a traffic source
- $c^2(S)$ = squared coefficient of variance of the silent (OFF) period of a traffic source
- $m(A)$ = mean active (ON) period of a traffic source, in cell emission times referenced to $PCR(VCC)$
- $m(S)$ = mean silent (OFF) period of a traffic source, in cell emission times referenced to $PCR(VCC)$
- b = burstiness factor of a traffic source = $\frac{PCR(VCC)}{SCR(VCC)} = \frac{m(A)+m(S)}{m(A)}$

Observe that any two of the last three parameters listed are sufficient. We elect to use $m(A)$ and b.

Consider now the functional dependencies of N^*, G and ρ_{out}. The value of N^* can be expressed as

$$N^* = N^*(\ B,\ K,\ CLR_{obj},\ MBS,\ c^2(A),\ c^2(S),\ m(A),\ b\ ) \tag{2}$$

and analogous dependencies exist for G and ρ_{out}. We wish to reduce the number of independent variables in (2). This reduction can be accomplished by fixing certain parameters as system constants, and by setting other parameters to values that yield the lowest, or

most pessimistic value, of N^*. The selection of model parameters to yield such a worst-case estimate is consistent with our objective of establishing a set of loading curves that can be used to provide conservative estimates of statistical multiplexing efficiency. It can be deduced from the relations provided in (1) that any selection of independent variable values to maximize N^* will also maximize G and ρ_{out}.

The independent variables B, CLR_{obj} and $m(A)$ can be fixed as constants. The value of B is set by the available multiplexing equipment. CLR_{obj} is a limiting value fixed by the operator of that multiplexing equipment, and is assumed here to be 10^{-7}. The value of $m(A)$ is set to approximate the mean Protocol Data Unit length associated with the type of traffic sources under consideration. Since these results are intended for validation against samples of Ethernet LAN traffic, $m(A)$ is taken here to be 28 cells†.

The independent variables MBS, $c^2(A)$ and $c^2(S)$ are selected to yield worst case values of N^*, as follows.

A number of considerations, including a preliminary set of simulation runs, indicate that N^* generally increases as MBS is made smaller. Therefore setting MBS to infinity results in the worst-case treatment of this independent variable. Since the source model based on a four-state traffic Markov Chain yields unbounded burst lengths, such worst case treatment is provided.

Preliminary simulation runs, as well as prior investigation (Eliazov (1990)), indicate that setting $c^2(A) = c^2(S) \equiv c^2$ has little affect on the resulting value of N^*. Hence this is done here. It remains to select appropriate values for c^2.

From simulation runs using representative parameter settings, the variation of N^* with respect to c^2 was observed (the details of these sample runs are omitted here due to space limitations). From such experiments, it was observed that after an initial decrease in N^*, a point exists beyond which further increases in c^2 do not yield a significant further decrease in N^*. The value of c^2 at this point is selected for establishing simulation-based estimates of N^*, G and ρ_{out}. Once this is accomplished, the functional dependency of the dependent variables upon c^2 is eliminated.

Simplified Functional Dependencies

After either fixing or selecting worst-case values for many of the independent variables in the manner described, N^* and the other dependent variables can be represented as functions of a reduced set of parameters. Hence functional dependencies of the type reflected by (2) can be replaced with

$$G = G(b, K); \quad N^* = N^*(b, K); \quad \rho_{out} = \rho_{out}(b, K) \tag{3}$$

With (3), these three measures of multiplexing efficiency are described in terms of the burstiness factor of the traffic sources and the capacity factor, which depends upon characteristics of both the traffic sources and the multiplexing equipment.

---

†The useful payload carried in 28 cells depends to some extent on the protocols used, but it is in the neighborhood of 1,300 octets. which is less than the 1,518 octet maximum length of an Ethernet packet.

4 SIMULATION ASPECTS AND LOADING CURVES

We now examine some further aspects of this simulation including characterization and treatment of cell loss events, and exhibit some representative loading curves for the initial phase of this methodology. Our simulator software is structured so that the simulated cell transfers (and cell loss events) associated with a single run are captured in a number of separately treatable "segments".

For one simulation run of about 10^{10} cell transfers, our simulator uses roughly 1,500 segments, each of which contains approximately 8×10^6 cell transfers. The observed CLR for one simulation run is calculated as the ratio of the number of cells lost to the total number of cells offered during that run.

On runs for which this observed CLR is in the neighborhood of 10^{-7} , it is found that all cell losses occur in a very few (typically less than 10) of these roughly 1,500 available segments. This is evidence of the tendency of cell losses to be clumped.

In our study, we make use of a "stopping rule" that permits the conclusion with reasonable confidence that $CLR_{obj} \equiv 10^{-7}$ is not exceeded on a particular simulation run, or on its indefinite extension. For a given set of independent variables, an efficient binary search procedure is used to select N for each simulation run, and to control convergence to N^* under the reasonable assumption that CLR is monotonically nondecreasing with N.

Loading Curves

We refer to a plot of N^* as a function of the capacity K (with all other independent variables being held constant) as a *loading curve*. Loading curves were generated for each of the three dependent variables over a range of burstiness factors and output buffer sizes. Other independent variables are either fixed or extremized as previously described. Figures 2, 3 and 4 show the set of loading curves obtained for N^*, G, and ρ_{out}, respectively, when $B = 3000$, and b takes selected values within a representative range.

5 ROLE OF A LAN TRACE-BASED SOURCE MODEL

Some measurement-based data for real traffic sources that might be statistically multiplexed on ATM networks are available in the form of suitably collected traces of Ethernet LAN traffic (Leland (1994), Erramilli (1994)). Such data are particularly useful for testing the applicability of our loading curves because such data have been shown to exhibit the self-similar characteristic that makes statistical multiplexing difficult.

Six such traces were examined in the second phase of this study, with each trace consisting of at least 10^6 Ethernet packets, and being equivalent to between 10^7 and 10^8 cells. The relevant information abstracted from such a trace are the time of a particular packet's occurence and the length of that packet. While the following results are framed for conciseness in terms of a single such trace, they generally apply to each of the Ethernet LAN traces that we have examined to date.

For the purpose of simulation using trace-based traffic sources, each source (or VCC as illustrated in Figure 1) effectively uses its own copy of the trace. To minimize time correlation across multiple sources, the starting poing within this trace is randomly and independently selected for each source. If and when the trace data is exhausted by any particular source, a second random starting point is selected. Because the random selection

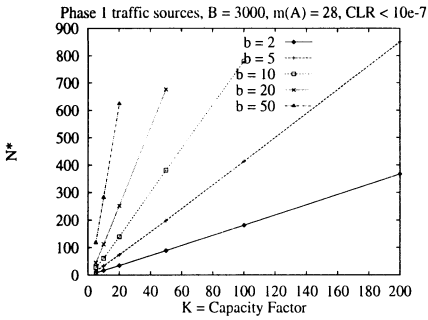


Figure 2:
Loading Curves for
Maximum Number of Traffic Sources

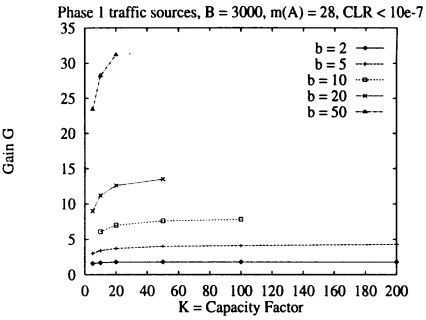


Figure 3:
Loading Curves for
Statistical Multiplexing Gain

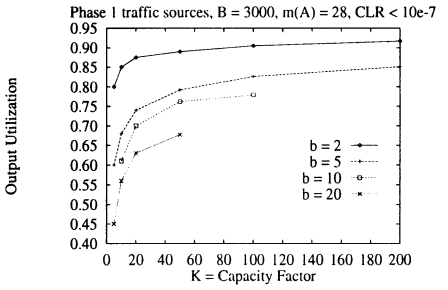


Figure 4:
Loading Curves for
Output Link Utilization

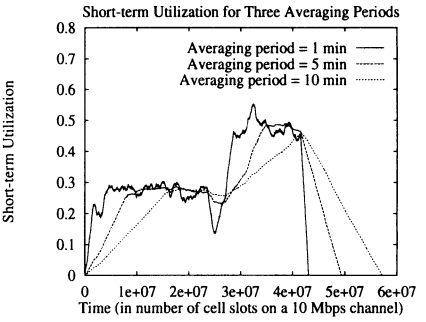


Figure 5:
Short-term Utilization of a Representative
Trace of LAN Traffic

of starting points is confined by our procedures to the initial portions of this trace, no source uses the trace more than twice when generating results reported here.

Some Statistical Properties of an Examined Ethernet Trace

As a preliminary step, the autocorrelation function for the basic trace was determined. This autocorrelation function indicates that some correlation does exist at the mean separation between the random starting times for the LAN trace copies used by different sources. We conclude from such correlation that the results of multiplexing simulations driven by these sources will understate to some extent the maximum number of such sources, N^*, achievable when operating in conformance with the loss objective, CLR_{obj}. This will result in somewhat conservative values for the three dependent variables, N^*, G and ρ_{out}, used here to characterize multiplexing efficiency. It also highlights, however, the

need to further pursue such investigations with considerably larger bases of real traffic data.

A number of statistics were obtained for the Ethernet trace under consideration after a preliminary conversion from packet-oriented data to cell-oriented data. This conversion is accomplished by treating each packet as a cell burst having suitable length and occurring at a $PCR(VCC)$ that is equivalent to 10 Mbits/second (for the Ethernet LAN). The average utilization, ρ_{source}^{long}, of this converted trace over its 30 minute duration is 0.33. The following measured values for this converted trace are noted:

- $m(A) = 24.8$ cells
- $m(S) = 49.7$ cells
- $c^2(A) = 4.63$
- $c^2(S) = 3.80$

Figure 5 displays three plots of the short term utilization for this converted trace that were obtained by averaging over sliding windows of 1 minute, 5 minutes and 10 minutes, respectively. A point on such a utilization plot is found through dividing the number of trace-originated cell slots occurring in a particular window (whose right end point is plotted on the abscissa) by the number of cell slots available in that window at the selected $PCR(VCC)$. The primary interest here concerns the central portions of these plots, with the end ramps being artifacts of data exhaustion in the sliding window averaging procedure. These plots of Figure 5 will be relevant when matching the multiplexing efficiencies achieved with trace-based sources to the efficiencies predicted by the loading curves given in Figures 2, 3 and 4.

Determining Safe Load Levels for Traffic Sources Based on Ethernet Traces

The next task is to determine safe traffic load levels, or equivalently, the maximum multiplexing efficiency that is safely achievable with a particular set of trace-based traffic sources. Except for the use of a different traffic source model, the maximum achievable multiplexing efficiencies for traffic sources based on one or more Ethernet traces are determined using methods similar to those presented in Section 4. For fixed values of all independent variables (e.g., $B = 3000$), the number of trace-based sources, N, is varied for each of a number of simulation runs. The largest admissible N, call it N^*, is next identified with a binary search procedure. Then knowing N^*, (1) is used to evaluate the remaining dependent variables, G and ρ_{out}.

Performing the above procedure for selected values of $K = 20$, 50 and 100 and plotting the results gives the desired safe load levels. These are shown for N^*, G and ρ_{out} as the solid curves in, respectively, Figures 6, 7 and 8.

Relation of Loading Curve Predictions to Ethernet Trace Results

Consider now the use of loading curves as provided in Figures 2, 3 and 4 for predicting the multiplexing efficiencies achievable with Ethernet-based traffic sources. The solid curves for N^*, G and ρ_{out} shown respectively in Figures 6, 7 and 8 are the results obtained with traffic sources modeled from an Ethernet trace. They are the results that one would wish to predict from these loading curves.

To apply these loading curves from the first study phase, it is necessary to identify the appropriate burstiness factor, b, to associate with the Ethernet-based traffic sources. The relation

$$b = \frac{1}{\rho_{source}} \tag{4}$$

implies that an appropriate value of b can be found from a properly selected value of ρ_{source}. But for most real traffic sources, ρ_{source} depends upon both the length of the

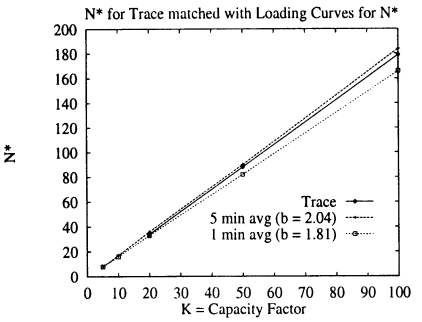


Figure 6:
Matching Trace and Loading Curve Results for Maximum Number of Traffic Sources

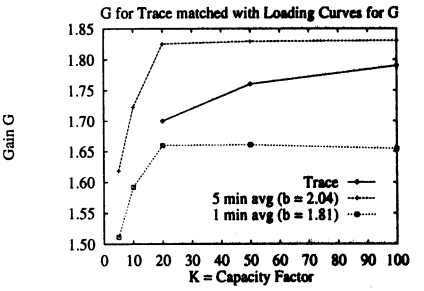


Figure 7:
Matching Trace and Loading Curve Results for Statistical Multiplexing Gain

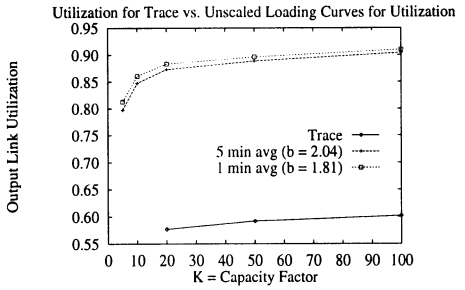


Figure 8:
Trace and *Unscaled* Loading Curve Results for Output Link Utilization

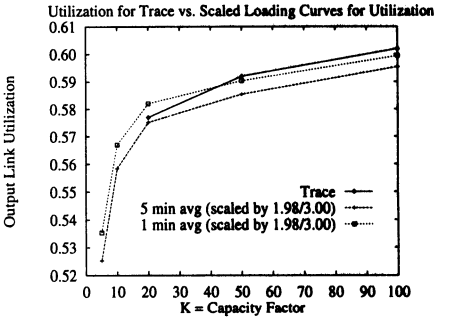


Figure 9:
Trace and *Scaled* Loading Curve Results for Output Link Utilization

averaging interval used to measure this quantity and the placement (or starting time) of that interval.

Several considerations are relevant here. First, for a given averaging interval length, a reasonable and conservative approach is to select the maximum value of ρ_{source} over an available period of observation. Second, the length of this averaging interval can be selected to yield a reasonable match between the results derived from a given set of data and the prediction methodology under test. The effectiveness of this prediction methodology can then be gauged by the degree to which this process can be reproduced with other sets of data and the same length of averaging interval.

By using plots of the type shown in Figure 5, maximum values of ρ_{source} can be identified for various averaging interval lengths. Then, based on the maximum value of ρ_{source} for

each plot, the corresponding value of b is determined with the help of (4). This value of b and a specified value of K can then be used to identify a corresponding point on a loading curve for N^*, G or ρ_{out}. This corresponding point is, of course, dependent upon the length of averaging interval used to generate the plot of ρ_{source}.

The execution of this process is demonstrated using the ρ_{source} plots from Figure 5 that correspond to the 1 minute and 5 minute averaging intervals. The resulting values of b are 2.04 for the 5 minute averaging interval and 1.81 for the 1 minute averaging interval. Then the initial loading curves in Figures 2, 3 and 4 are interpolated (or extrapolated) to yield the pairs of dashed curves shown in Figures 6, 7 and 8.

Comparing the 1 minute-based extrapolated loading curve with the trace-generated curve in Figure 6 shows a conservative prediction for N^*. A similar comparison for Figure 7 shows that a conservative prediction for G is achieved with the 1 minute-based extrapolated loading curve. Furthermore, we note that conservative predictions are found for N^* and G obtained from similar analyses using 1 minute averaging intervals for the remaining five Ethernet traces examined in this study.

In distinct contrast, the trace-generated curve for ρ_{out} shown in Figure 8 does not match either the 1 minute or 5 minute-based interpolated loading curves. This mismatch is not unexpected because related phenomena have been observed, whereby low order Markov Chain models significantly overstate the utilization levels of self-similar traffic that can be safely handled. The proper method for applying the initial loading curves for ρ_{out} is through an appropriate scaling procedure.

Scaling of the Loading Curve for Output Utilization

For steady state operation with negligible cell loss, it follows from continuity considerations that $N \cdot \rho_{source} \cdot PCR(VCC) = \rho_{out} \cdot PCR_{out}$. Setting N and ρ_{out} at their maximum permissible values yields

$$\rho_{source} = \frac{K}{N^*} \cdot \rho_{out} \tag{5}$$

Since ρ_{source} differs between the four-state Markov Chain model (and hence the loading curves) and the trace-based model, this relation indicates that ρ_{out} must differ proportionally as the source model changes. Interpolation of the b values shown in Figure 6 for the (dashed) 1 minute and 5 minute curves yields the corresponding value for this figure's solid curve as $b_{source} = 1.98$. As b_{source} has been matched to the loading curves – whose long term values of b are determinable from relations given in Section 3 – it is clear that ρ_{out} must be scaled by the ratio of the b_{source} to the long term value of b for the trace-based sources. Using the previously cited value $\rho_{source}^{long} = 0.33$ and (4) yields this quantity, $b_{source}^{long} = 3.00$.

Figure 9 replots ρ_{out} after this rescaling. The reasonable agreement shown here is also obtained for the other five Ethernet traces processed to date. We remark that for measurement-based applications, it appears possible to avoid the significant additional effort needed to evaluate b_{source} by approximating it with the reciprocal of the more easily measurable peak value of the appropriate short term utilization.

6 CONCLUSIONS AND EXTENSIONS

This paper describes a method for conservatively predicting the traffic carrying efficiency of ATM statistical multiplexers that are operating in compliance with a given cell loss objective. A suitable set of loading curves are first constructed using a traffic source model based upon a specific, low order Markov Chain. These loading curves are intended to support estimates of the highest multiplexing efficiency achievable when operating with real traffic sources. These loading curves are properly applied by using our identified scaling method, together with appropriate traffic measurements. The resulting predictions of multiplexing efficiency provided by this methodology have been shown, based on a limited amount of self-similar traffic data, to be conservative. This methodology's practical application is based upon a set of traffic measurements whose general availability may be reasonably anticipated, but which is not assured at present time. Additional refinement of this method is expected to further test its effectiveness on a larger base of traffic data and to better incorporate available traffic measurements.

REFERENCES

ANSI Std. T1.511-1994, *B-ISDN ATM Layer Cell Transfer - Performance Parameters.*

GR-1110-CORE (1995) *Broadband ISDN Switching System Generic Requirements*, Issue 1/Revision 2, (Bellcore).

Bensaou, B., Guibert, J., Roberts, J. W. (1990) Fluid Queuing Models for a Superposition of On/Off Sources. *7th ITC Specialist Seminar*, Paper 9.3.

Eliazov, T. E., Ramaswami, V., Willinger, W., Latouche, G. (1990) Performance of an ATM Switch: Simulation Study. *Proceedings of the IEEE Infocom '90*, San Francisco.

Elwalid, A., et al (1995) Fundamental Bounds and Approximations for ATM Multiplexers with Applications to Video Teleconferencing. *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 6, 1004–1016.

Erramilli, A., Gordon, J., Willinger, W., (1994) Applications of Fractals in Engineering for Realistic Traffic Processes. *Proceedings of the 14th International Teletraffic Congress*, 35–44, Amsterdam: Elsevier.

Fujimoto, R. M., Nikolaidis, I., Cooper, C. A. (1995) Parallel Simulation of Statistical Multiplexers. *Journal of Discrete Event Dynamic Systems — Theory and Applications*, Vol. 5. 115–140.

Hui, J. Y. (1990) Switching and Traffic Theory for Integrated Broadband Networks. Boston: Kluwer.

Draft Revised ITU Recommendation I.371 (1995) Traffic Control and Congestion Control in B-ISDN. ITU Study Group 13 Temporary Document Number 71(P), Geneva.

Leland, W. E., Willinger, W., Taqqu, M. S., Wilson, D. V. (1994) Statistical Analysis and Stochastic Modeling of Self-Similar Datatraffic. *Proceedings of the 14th International Teletraffic Congress - ITC14*, 319–328, Amsterdam: Elsevier.

Nikolaidis, I., Fujimoto, R. M., Cooper, C. A. (1993) Parallel Simulation of High-Speed Network Multiplexers. *32nd IEEE Conference on Decision and Control*, 2224–2229.

Nikolaidis, I., Fujimoto, R. M., Cooper, C. A. (1994) Time-Parallel Simulation of Cascaded Statistical Multiplexers. *ACM Sigmetrics Conference on Measurement & Modeling*, 231–240.