Models for Packet Switching of Variable-Bit-Rate Video Sources

PRODIP SEN, MEMBER, IEEE, BASIL MAGLARIS, MEMBER, IEEE, NASSER-EDDINE RIKLI, AND DIMITRIS ANASTASSIOU, MEMBER, IEEE

***Abstract*—Packet switching of variable-bit-rate real-time video sources is a means for efficient sharing of communication resources while maintaining uniform picture quality. Performance analysis for the statistical multiplexing of such video sources is required as a first step towards assessing the feasibility of packet video. This paper extends our earlier work in modeling video sources which have been coded using interframe coding schemes and in carrying out buffer queueing analysis for the multiplexing of several such sources. Our previous models and analysis were suitable for relatively uniform activity scenes. Here, we consider models for scenes with multiple activity levels, which lead to sudden changes in the coder output bit rates. We present correlated Markov models for the corresponding sources, and using a flow-equivalent queueing analysis, obtain common buffer queue distributions and probabilities of packet loss. Our results demonstrate the efficiency of packet video on a single link, due to the smoothing effect of multiplexing several variable-bit-rate video sources.**

I. Introduction

PACKETIZED transmission in an asynchronous transfer mode (ATM) ISDN decouples the user input from the network by providing a unified transport mechanism for services of widely varying baud rates. In addition, it can perform statistical multiplexing by taking advantage of statistical variations in the traffic offered by users. In video communications, variable-bit-rate compression algorithms transmit at a higher rate during high-activity (motion) scenes and at a low rate when there is less motion. It is possible to multiplex statistically several independent video transmissions at a speed lower than the aggregate peak coding rate. The law of large numbers indicates that as the number of independent sources increases, the aggregate rate approaches the average, without adjustment of individual source rates by varying the picture quality. Equivalently, the probability of buffering or delaying data beyond a certain threshold decreases. This probability is related to the fraction of packets that arrive at their destination in time to be played back; thus, it is a major performance index. We develop queueing models to assess this probability.

The techniques and results summarized below are extensions of previous results reported in [1]. They were motivated by experimental data obtained at Bell Communications Research. In that earlier paper, we presented correlated Markov process models for video sources coded using conditional replenishment interframe coding. The models were applicable to video scenes with relatively uniform activity levels, such as scenes showing a talking person. In what follows, we extend these models to encompass simultaneous multiplexing of two kinds of scenes: slow varying and fast varying. Such models apply to talker–listener alternating scenes, as well as to situations where there is a mix of dissimilar services, e.g., television and videotelephony.

II. The Video Source Model

We consider digital video sources which are compressed using interframe variable-rate coding [2]. The coded bit stream from each source is stored in a separate prebuffer, which assembles the data into blocks (typically a frame's worth of data) and packetizes the blocks. Prebuffering eliminates complicated properties in the nature of the source model [1], [3]. The packets from all the prebuffers join a common buffer in the multiplexer, where the packets are queued for transmission over a high-speed communication line. The schematic setup is shown in Fig. 1 [3].

For the situation we consider, the data rates will be on the order of megabits per second, while the packet lengths will be less than a kilobit. Thus, it is possible to ignore the discrete packet nature of the data and treat them as a continuous bit stream or flow. As a result, we model the sources as producing continuous bit streams at quantized data-rate levels, with probabilistic transitions between the various rate levels. Correspondingly, we also model the statistical multiplexer queue as a fluid-flow pipe which takes in bits from the various prebuffers and serves them at a constant rate. The fluid-flow approximation is a pow-

Manuscript received October 26, 1988. This work was supported in part by the New York State Science and Technology Foundation, through its Center for Advanced Technology in Telecommunications, Polytechnic University, Brooklyn, NY; in part by the New York State Center for Advanced Technology in Computers and Information Systems and the Center for Telecommunications Research, Columbia University, New York; and in part by the National Science Foundation under Grant PYI:ECS-84-51499. This work was presented in part at the SPIE Conference on Visual Communications and Image Processing, Cambridge, MA, Oct. 1987.

P. Sen, B. Maglaris, and N.-E. Rikli are with Polytechnic University, Brooklyn, NY 11201.

D. Anastassiou is with the Department of Electrical Engineering, Columbia University, New York, NY 10027.

IEEE Log Number 8927597.

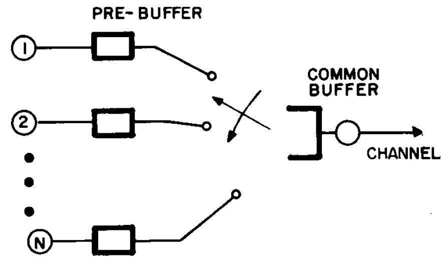


Fig. 1. Schematic of a statistical multiplexer.

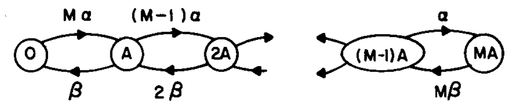


Fig. 2. State-transition-rate diagram for a single-activity-level source model.

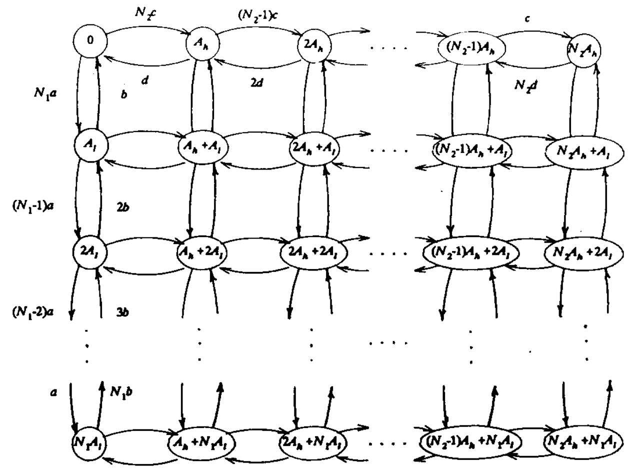


Fig. 3. State-transition-rate diagram for the aggregate source model.

erful tool which allows the use of analytic models, taking into account the source correlations in the queueing analysis.

Our earlier model, presented in [1], models the aggregated outputs of all the coders as a correlated Markov process whose state-transition-rate diagram is shown in Fig. 2. The aggregated process can transit between $N + 1$ data rate levels where level i corresponds to data rate iA. The number of quantization levels N was chosen arbitrarily, while the rate increment A and the transition rates α and β were chosen to match the mean, variance, and autocovariance function of the experimental data. The appropriateness of this Markov model stems from the data in [1] and from earlier work [3], [4], as well as from some more recent experiments [5]. These results indicate that an exponential correlation model for the data-rate process is a very good aproximation for videotelephone scenes with a uniform activity level, e.g., showing a person talking.

For other types of video traffic, such as broadcast television, videoconferencing, and longer videotelephone sequences (showing persons talking and listening), experimental work indicates the following structure. If we consider an environment where the video sources feeding the network are a mix of these types, then two important correlations are evident: a relatively fast-decaying short-term correlation corresponding to uniform activity levels, with a time constant on the order of a few hundred milliseconds, and a slow-decaying long-term correlation corresponding to sudden changes in the gross activity level of the scene (e.g., scene changes in broadcast TV or changes between listener and talker modes in a videotelephone conversation), with a time constant on the order of a few seconds [6]. Our earlier model [1] captured only the short-term correlation.

In this paper, we extend our model to accommodate the above-mentioned correlation structure. Moreover, we allow the multiplexing of *statistically different* sources, with different means and variances for the bit rates. We approximate the correlation decays as exponential since we feel that this captures the essential features of the correlation and provides a model which lends itself to analysis. The time constants for the two decay rates are different, in general, and are matched to the data.

Our extended model, which includes both short-term and long-term correlations, involves a correlated Markov process model with a state-transition-rate diagram as shown in Fig. 3. Our models represent the source as one which changes among different fixed-rate levels. The label in any state indicates the data rate out of the prebuffer corresponding to that state. The possible data-rate levels are built up from two basic levels: a high rate A_h and a low rate A_l, via integer combinations up to a maximum of $N_1 + 1$ low-rate levels and $N_2 + 1$ high-rate levels. Note that this generalized model handles abrupt changes in the output rate, unlike the earlier one. In the special case of a videotelephone sequence, where a person alternates between talking and listening, the individual source model will reduce to the case of $N_2 = 1$, as shown in Fig. 4.

When several sources are multiplexed, the resultant *aggregate* bit rate can be modeled by the *same structure* as the individual source model. The sources need not be statistically identical: they may have different means and variances. We assume only that the autocovariance behavior of all of the sources can be approximated by the *same* two dominant time constants (a "fast" mode and a "slow" mode). To determine the rest of the parameters in the model, first- and second-order statistics are matched. The maximum rates can also be equated.

As an example, consider a single videotelephone source, involving transition between talking and listening (the model of Fig. 4). In this case, the fraction of time spent in the high activity level and the average time spent in the high level are used to fix c and d. The ratio of the average data rate in the high activity level to that in the low activity level is defined as the *mean ratio* γ. Matching the mean ratio, the overall mean $(\overline{\lambda})$, and the second-

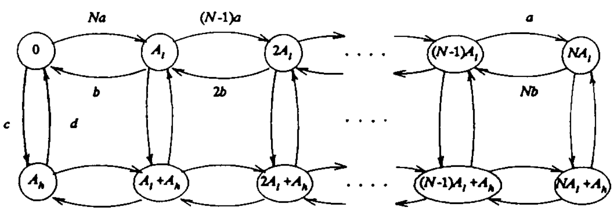


Fig. 4. State-transition-rate diagram for a two-activity-level source model.

order statistics in a single activity level [the conditional autocovariance function $C(\tau)$] *completely determines* the values of all of the other parameters, a, b, A_l, and A_h. The parameter N indicating the number of quantization levels in any activity level is the only free parameter to be chosen as desired. The equations to be used for the matching are

$$C(\tau) = C(0)e^{-(a+b)\tau} \tag{1}$$

$$C(0) = Np(1-p)A_l^2 \qquad \text{where } p = \frac{a}{a+b} \tag{2}$$

$$\gamma = \frac{NpA_l + A_h}{NpA_l} \tag{3}$$

$$\bar{\lambda} = NpA_l + qA_h \qquad \text{where} \quad q = \frac{c}{c+d}. \tag{4}$$

The specific order to determine the parameters of the model is as follows. From the actual data, the fraction of time spent in the high activity level can be equated to q, and the average time spent in the high level can be equated to $1/d$. This fixes both c and d. Matching the conditional variance, the conditional autocovariance exponent, the mean ratio, and the overall mean, with the help of (1)–(4), yields the values of the parameters a, b, A_l, and A_h. When M such sources are multiplexed, the overall process can then be represented by our general model (Fig. 3), with the above parameters, and $N_2 = M$, $N_1 = MN$, where N is a freely chosen parameter.

III. Performance Analysis

In this section, we describe the analysis of the multiplexer queue using the source model of the previous section, considering the queue to be of constant fluid flow. The process of Fig. 3 can be decomposed into a superposition of simpler sources. In fact, the process is just a superposition of independent *ON–OFF miniprocesses*, N_1 of the type shown in Fig. 5(a) and N_2 of the type shown in Fig. 5(b). The aggregate source process state then corresponds to the pair (i, j), denoting the respective number of miniprocesses which are ON.

Let μ be the fixed output rate of the continuous state multiplexer queue and $q(t)$ be the instantaneous queue size. The process of Fig. 3 feeds the multiplexer queue. By considering the Chapman–Kolmogorov forward equations for the joint probability distribution of source state

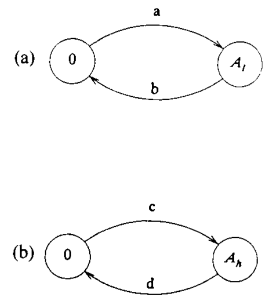


Fig. 5. Miniprocess models.

and multiplexer queue size, at steady state, the following equation can be obtained

$$\begin{aligned}\frac{dF_{i,j}(x)}{dx} &= \frac{(N_1 - i + 1)a}{\lambda_{i,j} - \mu} F_{i-1,j}(x) \\ &\quad - \frac{ib + jd + (N_1 - i)a + (N_2 - j)c}{\lambda_{i,j} - \mu} F_{i,j}(x) \\ &\quad + \frac{(i+1)b}{\lambda_{i,j} - \mu} F_{i+1,j}(x) \\ &\quad + \frac{(N_2 - j + 1)c}{\lambda_{i,j} - \mu} F_{i,j-1}(x) \\ &\quad + \frac{(j+1)d}{\lambda_{i,j} - \mu} F_{i,j+1}(x)\end{aligned} \tag{5}$$

where

$$F_{ij}(x) = \text{Prob(source is in state } (i, j), \text{ multiplexer queue size} \le x) \tag{6}$$

and $\lambda_{ij} = iA_l + jA_h$. Equation (5) can be written as

$$\boldsymbol{D\dot{F}} = \boldsymbol{MF} \tag{7}$$

in which $\boldsymbol{D}$ and $\boldsymbol{M}$ are appropriate matrices and $\boldsymbol{F}$ is the vector formed from the F_{ij}.

The solution of (7) is given by

$$\boldsymbol{F}(x) = \boldsymbol{F}(\infty) + \sum_z a_z \boldsymbol{\phi}^{(z)} e^{zx} \tag{8}$$

where the z are eigenvalues of $\boldsymbol{D}^{-1}\boldsymbol{M}$ in the left half complex plane and $\boldsymbol{\phi}^{(z)}$ are the corresponding eigenvectors

$$z\boldsymbol{D\phi}^{(z)} = \boldsymbol{M\phi}^{(z)}. \tag{9}$$

Note that for the solution of (7) to be a probability distribution function, only the left half plane eigenvalues can appear.

To obtain the complete queue distribution, we need to evaluate the eigenvalues and eigenvectors of $\boldsymbol{D}^{-1}\boldsymbol{M}$, as well as the coefficients a_z. We first determine the eigenvector for a given eigenvalue. Let $\Phi^{(z)}(u, v)$ denote the generating function of $\phi^{(z)}$

$$\Phi^{(z)}(u, v) = \sum_{i=0}^{i=N_1} \sum_{j=0}^{j=N_2} \phi_{ij}^{(z)} u^i v^j. \tag{10}$$

Using (10), the following equation for the generating function can be obtained (we omit the dependence on z for notational clarity)

$$\frac{\partial \Phi(u, v)}{\partial u}\left[-b + (A_l z + b - a)u + au^2\right] + \frac{\partial \Phi(u, v)}{\partial v}\left[-d + (A_h z + d - c)v + cv^2\right] = \Phi(u, v)\left[\mu z - (N_1 a + N_2 c) + N_1 a u + N_2 c v\right]. \tag{11}$$

We solve (11) by using a separation of variables. Thus, letting

$$\Phi(u, v) = h(u) \cdot g(v), \tag{12}$$

substituting in (11) we obtain

$$h(u) = (u - r_1)^{c_1}(u - r_2)^{c_2} \tag{13}$$

$$g(v) = (v - r_3)^{c_3}(v - r_4)^{c_4} \tag{14}$$

where

$$r_1 = \left\{-(A_l z + b - a) + \left[(A_l z + b - a)^2 + 4ab\right]^{1/2}\right\}/2a$$

$$r_2 = \left\{-(A_l z + b - a) - \left[(A_l z + b - a)^2 + 4ab\right]^{1/2}\right\}/2a$$

$$r_3 = \left\{-(A_h z + d - c) + \left[(A_h z + d - c)^2 + 4cd\right]^{1/2}\right\}/2c$$

$$r_4 = \left\{-(A_h z + d - c) - \left[(A_h z + d - c)^2 + 4cd\right]^{1/2}\right\}/2c$$

and

$$c_1 = \frac{\mu z/2 + \Gamma - N_1 a(1 - r_1)}{a(r_1 - r_2)}; \qquad c_2 = N_1 - c_1$$

$$c_3 = \frac{\mu z/2 - \Gamma - N_2 c(1 - r_3)}{c(r_3 - r_4)}; \qquad c_4 = N_2 - c_3.$$

The eigenvector generating function $\Phi(u, v)$ can be obtained from (12)–(14), and the eigenvector components can be obtained by comparing this expression to the expansion of (10). Given an eigenvalue z, the complete procedure for finding the eigenvector is the following.

Compute $r_1 - r_4$ and c_1, c_3 from the above, with Γ chosen to make both c_1 and c_3 integers. Use (12)–(14) to obtain Φ and (10) to obtain the eigenvector components.

To obtain the eigenvalues, we proceed as follows. Solving for Γ in the expressions for c_1 and c_3 above and squaring both sides of the resulting equation, we get the fourth-order equation

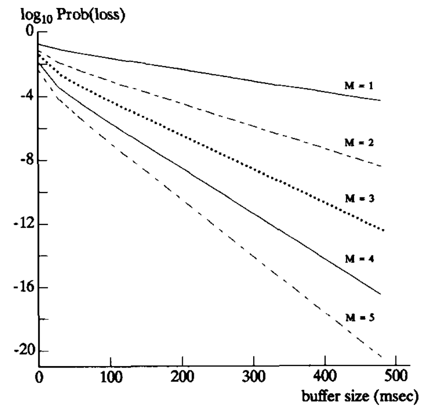


Fig. 6. Variation of loss probability with buffer size for a utilization of 65 percent and a mean ratio of 1.5.

$$T_0 z^4 + T_1 z^3 + T_2 z^2 + T_3 z + T_4 = 0 \tag{15}$$

where the coefficients T_0, T_1, T_2, T_3, and T_4 are all functions of *only* the model parameters. We omit their explicit expressions. The eigenvalues are obtained by solving (15) for all possible combinations of c_1 and c_3.

Once the eigenvalues and the eigenvectors are obtained as above, the coefficients a_z of (8) have to be evaluated in order to determine the complete queue distribution. To that end, we use the fact that the queue size is nonzero with probability one, if the instantaneous input rate is greater than the queue output rate, since we have a flow model for the queue. This gives rise to a set of linear equations

$$F_{ij}(0) = 0 \quad \text{if} \quad \lambda_{ij} = (iA_l + jA_h) > \mu. \tag{16}$$

The buffer overflow probability, or "survivor function," for a given queue size x is

$$G(x) = \text{Prob}(\text{queue size} > x) = 1 - \sum_{n,m} F_{nm}(x). \tag{17}$$

IV. Results

In this section, we present some results generated by our analysis. Our intent is to show the types of results which can be obtained, as well as to note general trends. We present results for the special case of a videotelephone conversation. Thus, each source is modeled as in Fig. 4. The mean data rate per video source is taken to be 3.9 Mbits/s, the source-data-rate conditional variance is 3.015 Mbits^2/s^2, and the short-term correlation exponent is 3.9/s. For the long-term correlation parameters, we choose $c = d$. This is motivated by the videotelephone application we have in mind, where a scene alternates between a person talking and listening for approximately equal lengths of time on the average. The average time spent in any one state is taken to be 1.5 s. The queue

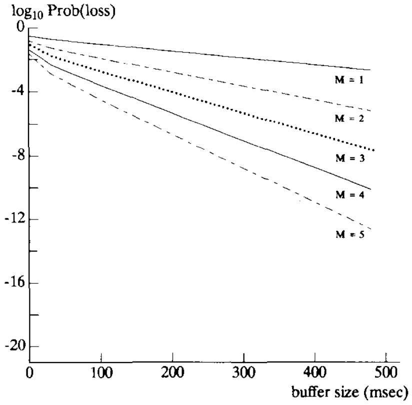


Fig. 7. Variation of loss probability with buffer size for a utilization of 65 percent and a mean ratio of 2.5.

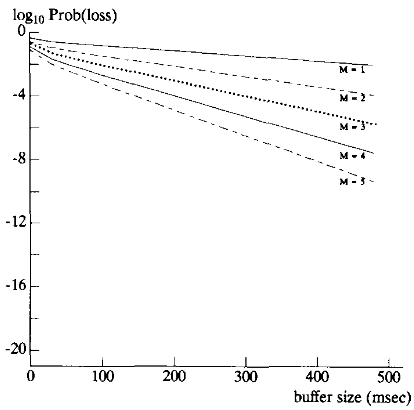


Fig. 8. Variation of loss probability with buffer size for a utilization of 75 percent and a mean ratio of 1.5.

output rate is adjusted according to the chosen utilization. The mean ratio is also varied.

In Figs. 6–8, we show the survivor function for the queue size in the multiplexer queue for various combinations of utilization values and mean ratio. The ordinate $G(x)$ thus shows the probability of data loss if the buffer size exceeds x. The buffer size is specified in time units, representing the time required to empty a full buffer at the output rate. We choose combinations of low and high utilizations and low and high mean ratio. Each graph shows the loss probability for multiplexing one to five video sources, demonstrating the dramatic reduction in loss probability as the number of multiplexed sources increases. In order to achieve the same loss probability with higher utilization (compare Figs. 6 and 8) or higher mean ratio (compare Figs. 6 and 7), a larger number of sources have to be multiplexed.

References

[1] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, pp. 834–844, July 1988.

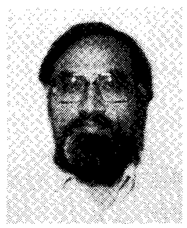
[2] A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression.* New York: Plenum, 1989.

[3] B. G. Haskell, "Buffer and channel sharing by several interframe picturephone coders," *Bell Syst. Tech. J.*, vol. 51, no. 1, pp. 261–289, Jan. 1972.

[4] J. O. Limb, "Buffering of data generated by the coding of moving images," *Bell Syst. Tech. J.*, vol. 51, no. 1, pp. 239–255, Jan. 1972.

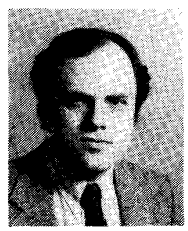
[5] W. Verbiest, "Video coding in ATD environment," in *Proc. Third Int. Conf. New Syst. Services Telecommun.*, Liege, Belgium, Nov. 1986.

[6] ——, "The impact of ATM concept on video coding," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 1623–1632, Dec. 1988.



Prodip Sen (S'75–M'81) was born in Amritsar, India, in 1951. He received the B.Tech. degree in electrical engineering from The Indian Institute of Technology, Bombay, India, in 1973 and the Ph.D. degree in electrical engineering from the Indian Institute of Science, Bangalore, India, in December 1977.

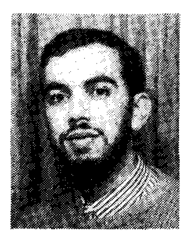
He was a Research Fellow and Visiting Assistant Professor in the Department of System Science in the School of Engineering at the University of California, Los Angeles, from 1978 to 1980. From 1980 to 1983 he was a Research Mathematician at Polysystems Analysis Corp., Huntington, NY. Since 1983 he has been with Polytechnic University, Brooklyn, NY, where he is currently an Associate Professor of Electrical Engineering. His research interests are in the areas of integrated networks, local and metropolitan area networks, and network management.



Basil Maglaris (S'74–M'79) was born in Athens, Greece, in 1952. He received the undergraduate Diploma degree in electrical engineering from the National Technical University of Athens in 1974, the M.Sc. degree from the Polytechnic Institute of Brooklyn, Brooklyn, NY, in 1975, and the Ph.D. degree in electrical engineering and computer science from Columbia University, New York, in 1979.

From 1979 to 1981 he was with the Network Analysis Corporation, Great Neck, NY, where he was involved in several projects in data and voice networks for both government and industry. In 1981 he joined the Polytechnic Institute of New York, Brooklyn, where he is currently an Associate Professor of Electrical Engineering and Computer Science. His research interests focus on the analysis, performance evaluation, and optimization of data, voice and integrated networks, packet radio, and local area networks.

Dr. Maglaris has been involved in various professional activities with the IEEE and the ACM.



Nasser-Eddine Rikli received the Diplome d'Ingenieur d'Etat degree (with honors) in "genie electrique" from the Institut National d'Electricite et d'Electronic, Boumerdes, Algeria, in 1984 and the M.S. and Ph.D. degrees in electrical engineering from Polytechnic University, Brooklyn, NY, in 1985 and 1988, respectively.

His main interests are in the modeling and analysis of integrated communication networks involving voice, data, and video and mobile communication with application to cellular radio.

Dimitris Anastassiou (S'77–M'78), for a photograph and biography, see this issue, p. 631.