Methods for Performance Evaluation of VBR Video Traffic Models

David M. Lucantoni, Marcel F. Neuts, *Member, IEEE*, and Amy R. Reibman, *Member, IEEE*

***Abstract*— Models for predicting the performance of multiplexed variable bit rate video sources are important for engineering a network. However, models of a single source are also important for parameter negotiations and call admittance algorithms. In this paper we propose to model a single video source as a Markov renewal process whose states represent different bit rates.**

We also propose two novel goodness-of-fit metrics which are directly related to the specific performance aspects that we want to predict from the model. The first is a leaky bucket contour plot which can be used to quantify the burstiness of any traffic type. The second measure applies only to video traffic and measures how well the model can predict the compressed video quality.

I. Introduction

IT is well recognized that the viability of B-ISDN/ATM depends on the development of effective and implementable congestion control schemes. While many frameworks and techniques are under discussion (see, e.g., [1]), at least two capabilities have been agreed to as necessary in any framework that might arise.) The first is a connection admission control (CAC) by which the network will decide to accept or reject a new connection based on a set of agreed to traffic descriptors and on available resources. Once a connection is accepted, a second necessary control is some form of usage parameter control (UPC) which will insure that connections stay within their negotiated resource parameters. A popular UPC would involve a leaky bucket monitor of traffic entering the system, where traffic deemed as excessive by the monitor could either be dropped or tagged as low priority and allowed to proceed through the network to take advantage of potentially unused resources.

Performance modeling is necessary to determine which techniques or set of techniques will be appropriate for eventual implementation in a B-ISDN network. Such models need to take into account traffic characteristics from realistic services that would be carried in a B-ISDN network. In particular, we need traffic models which will accurately represent the statistical nature of very high-speed, bursty services.

Two classes of traffic models need to be developed: multiplexed source models and single source models. Although the same traffic model might be used in both cases, some models might be more suitable for one than the other. Multiplexed models will capture the effects of statistically multiplexing bursty sources and will predict to what extent the superposition of bursty streams is "smoothed". These models will be useful in traffic engineering the network (e.g., deciding how many links or virtual paths to put between different locations) and in traffic management (e.g., designing connection admission control algorithms, etc.) Several models have already been proposed in this direction (see, e.g., [2], [3], [4], [5] and the references there).

There are several areas where single source models are useful. They could be used to study what types of traffic descriptors make sense for parameter negotiation with the network at call setup. For example, if leaky bucket monitoring is used as a traffic descriptor, the negotiation might consist of the source specifying what parameters could be used in the leaky bucket for a given connection. Single source models can help in the selection of these parameters. Also, some applications may do some end-to-end rate control to ensure that minimal traffic is lost during periods of network congestion. Source models could be used in testing various rate control algorithms. Finally, these models are also useful in predicting the quality-of-service (QOS) that a particular application might experience during different levels of congestion.

In deriving traffic models, we need metrics which can determine how "close" the model is to the actual traffic. Standard statistical measures such as means, variances, and other goodness-of-fit tests may not be appropriate here since they may not be measuring the characteristics of the process that are most important for either predicting the effect of the source on the resources in the network or the performance the source will experience. Instead, the goodness-of-fit metrics need to be directly related to the specific aspects of performance that we want to predict from the model; see e.g., [6].

In this paper, we propose two criteria for judging the appropriateness of a traffic model for bursty services. The first one applies to any high speed bursty data service and the second is specific to a variable-bit-rate (VBR) video application. To illustrate these measures we compare a previous model of VBR video with a new model proposed here.

II. Modeling Variable-Bit-Rate Video

The data we are modeling was recorded at an actual teleconference meeting. Each scene depicts the head and shoulders of one person, and is 5 min, or 9000 frames, long. Since each 5 min of video required approximately one week to encode using software, the motivation for developing accurate models with a low computational burden is clear. A typical

Manuscript received November 1992; revised February 1994; approved by IEEE/ACM Transactions on Networking Editor D. Raychaudhuri.

D. M. Lucantoni and A. R. Reibman are with AT&T Bell Laboratories, Holmdel, NJ 07733 USA.

M. F. Neuts is with the Department of Systems and Industrial Engineering, University of Arizona, Tucson, AZ 85721 USA.

IEEE Log Number 9402100.

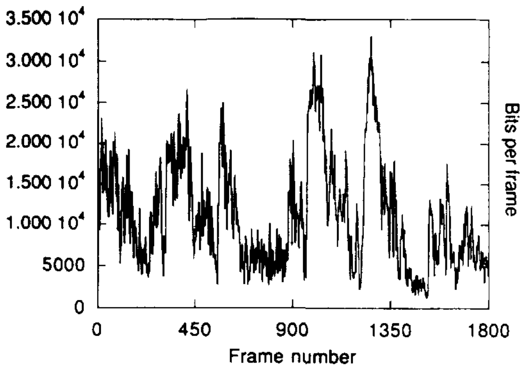


Fig. 1. Original data.

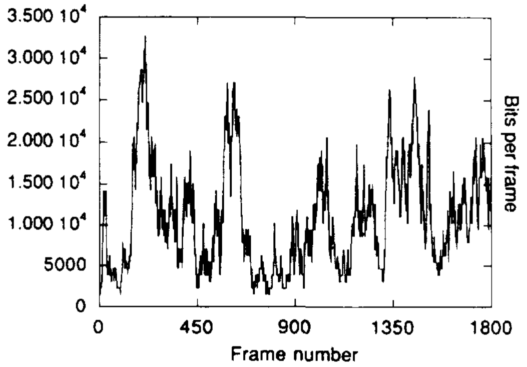


Fig. 2. MRP model.

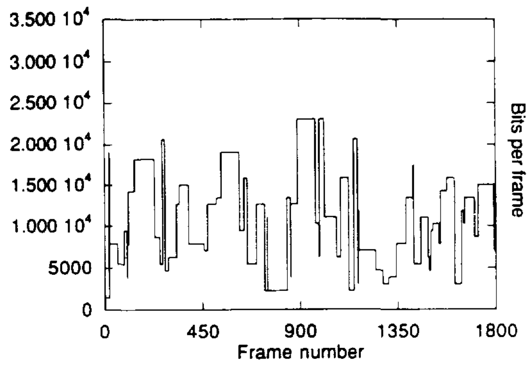


Fig. 3. DAR model.

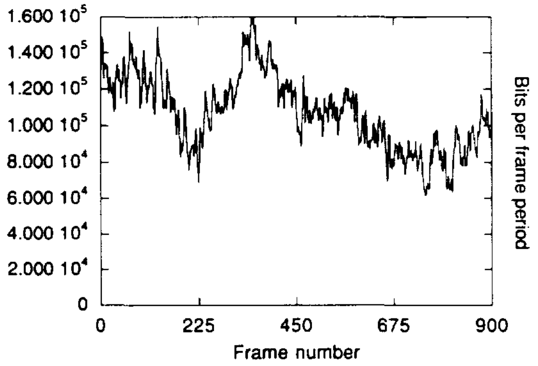


Fig. 4. Original data, 10 sources.

sample path is shown in Fig. 1, which shows the number of bits for 1800 successive frames or one minute.

We now briefly present a Markov renewal process model [7] of VBR video and review a recently proposed discrete autoregressive model.

A. A Markov Renewal Process (MRP) Video Model

We partition the range of possible rates (i.e., bits per frame) into 40 equidistant levels. The transition matrix P is estimated empirically. Next, we estimate the sojourn time distribution in each level. We fit the empirical distributions to a mixture of two geometric distributions, when possible, and otherwise we fit them to geometric distributions. Fitting all of the sojourn time distributions to geometric distributions, while giving satisfactory results, did not perform as well as the mixtures. Sampling from the fit distributions actually produced better results than the original empirical distributions.

A sample of 1800 frames generated by the MRP model is shown in Fig. 2. This certainly "looks" somewhat like the original data shown in Fig. 1 but required much less time to generate than the coded traffic. However, in Sections III and IV we use two potential measures to determine how accurately this model predicts actual source performance.

B. A Discrete Autoregressive Model (DAR)

A discrete autoregressive (DAR) model has recently been proposed for VBR video traffic by Heyman *et al.*, [5] and has been shown to predict accurately the blocking characteristics of a superposition of video sources into a statistical multiplexer. This is a very simple 3-parameter model but compares favorably to other models when predicting multiplexed traffic. We chose this model to compare with the MRP model because of its simplicity and to help illustrate the measures to be described below.

As discussed in Heyman *et al.*, [5], the DAR model is a Markov chain with transition matrix

$$R = \rho I + (1 - \rho)Q, \qquad (1)$$

where ρ is the autocorrelation and Q is a matrix with identical rows equal to the marginal distribution. The form of R in (1) gives some insight into the behavior of the model. Since the autocorrelation is on the diagonal, for high values (we obtained $\rho \approx 0.98$ in our data) the Markov chain will stay in a state for a long time. When it leaves, it chooses the next state according to the marginal distribution, so that it tends toward the mean. Although it stays in a high-rate level for a fairly long time, it may not stay in a group of high levels as long as it should and therefore does not capture the burstiness sufficiently.

Fig. 3 shows a sample of 1800 frames generated from the DAR model. This confirms the above observation that the sojourn times in given levels can be long. While the sample path doesn't "look" like the original data, the overall trends are close and the model may still work well in predicting the specific performance measures of interest. In particular, Figs. 4-6, show plots from the data, MRP model, and DAR model respectively, from the superposition of 10 identical video sources. Both models "look" like the data in this case.

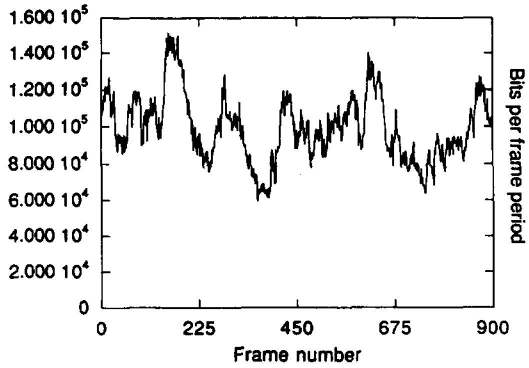


Fig. 5. MPR model, 10 sources.

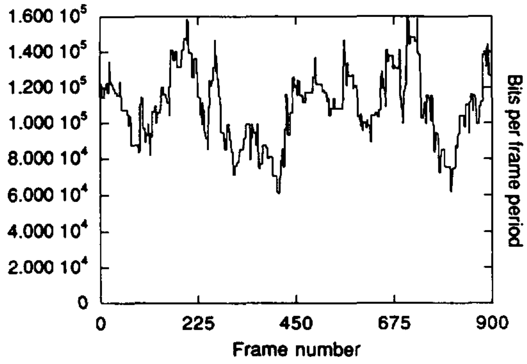


Fig. 6. DAR model, 10 sources.

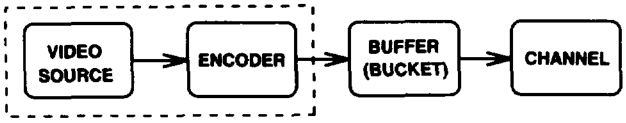
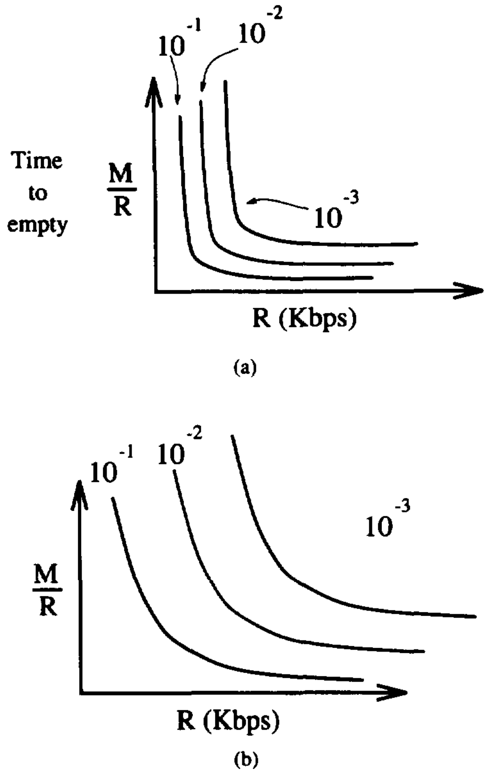


Fig. 7. A video system.

III. Leaky Bucket Contours

A diagram of a generic video system is shown in Fig. 7. A buffer/bucket lies between the video encoder (which generates the bits per frame) and the transmission channel. This could represent either a physical buffer that contains the actual traffic and performs some type of traffic shaping, or a logical buffer such as a leaky bucket counter [8] which only monitors and does not buffer the traffic. CBR video systems typically use the former to generate a constant bit rate onto the channel, while VBR systems could use the latter. In either case the values for the buffer parameters are needed to ensure that a given percentage of cells are not excessive. These parameters are directly related to the average rate and burstiness of the source. Since a physical buffer and a leaky bucket are logically equivalent for a given drain rate, we focus our attention on leaky buckets.

It has been proposed previously (see, e.g., [8]) that the leaky bucket parameters (e.g., bucket size and drain rate) could be passed to the network as a descriptor of the traffic characteristics. However, to judge the goodness-of-fit of a model or to compare different models we would like to have a more comprehensive characterization of the traffic. We propose the concept of *leaky bucket contours* as a way to describe the traffic characteristics of a bursty data source.



(a)

(b)

Fig. 8. A characterization of traffic burstiness.

Typical plots of leaky bucket contours are shown in Fig. 8. The drain rate of the leaky bucket is plotted on the x-axis and the size of the bucket is plotted on the y-axis. Points along the different curves are those pairs of (drain rate, bucket size) which result in a fixed percentage of traffic overflowing a leaky bucket with the given parameters. If we think of representing the probability of overflow as the z-axis coming out of the page, then the curves that are shown can be viewed as contour lines of a 3-dimensional surface which is of height 1 at the origin of the $x-y$ axes and which decreases in height as we get further into the upper right hand quadrant. The steepness with which this surface approaches zero on the z-axis is directly related to the burstiness of the traffic.

These leaky bucket contours capture the burstiness on many time scales simultaneously. A very good feeling for the burstiness of the source can be obtained by observing the gradients along the surface at these higher probability contours. The contour plots shown in Fig. 8(a) are clearly from a less bursty source than that shown in Fig. 8(b). At least one of these contours may help the end system to decide which parameter values to choose as a description of its traffic to the network.

We have shown typical contours for overflow probabilities $10^{-i}, i = 1, 2, 3$. At first this may not seem consistent with the view that typical leaky bucket parameters might be chosen to result in overflow probabilities of $10^{-6} - 10^{-9}$ or less. However, we believe that the higher probability contours are particularly relevant.

There are two options for a data source that wants to ensure that no more than 10^{-9} percentage of its traffic would overflow

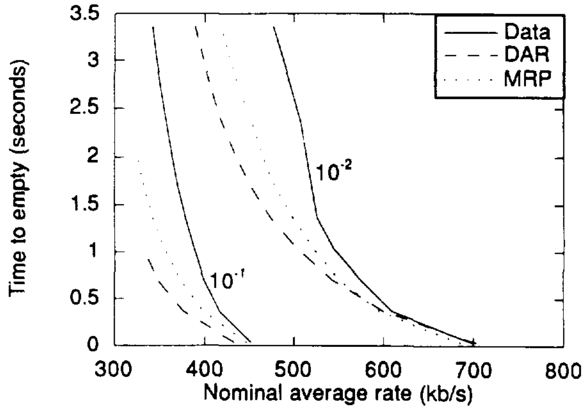


Fig. 9. Leaky bucket size for given overflow probability.

a leaky bucket. The first is to select a very large drain rate and/or bucket size for the traffic description. This would result in the network "reserving" a large amount of resources for the source and charging a high price to the end user. The second option, which we believe is more likely for most sources, is to reduce the cost of transport by selecting more reasonable leaky bucket parameters and to do at least a minimal amount of traffic shaping to ensure that no more than the desired percentage of traffic would overflow the bucket. Most sources would probably set the parameters in the $10^{-1} - 10^{-3}$ range and shape the traffic when necessary.

Unfortunately, these are also the only contours that can be accurately estimated with the limited amounts of data available. For example, we required somewhere between 10^4 and 10^5 frames to get an accurate estimate of the 10^{-2} curve, since these are not *independent* samples but are in fact highly correlated. Therefore, curves for 10^{-6} and lower cannot be accurately estimated for any reasonable simulation run.

Fig. 9 shows the 10^{-1} and 10^{-2} contour plots for the actual VBR video traffic as well as the MRP and DAR models both using 40 equidistant levels. The "time to empty" on the y axis is defined as M/R where M is the leaky bucket size and R is the drain rate of the bucket. Both models predict the lower bucket sizes fairly well, but as the drain rates decrease and bucket sizes increase both models become less accurate. The MRP model seems to outperform the DAR model over the entire range. One possible explanation for this is that as mentioned earlier, the behavior of the DAR model inherently underestimates the burstiness since it doesn't stay in a group of high states over any interval as long as the original process does.

IV. Quantization Histograms

The leaky bucket contours discussed in the last section can be used to compare the burstiness of different sources or to quantify the goodness-of-fit of a traffic model to a source. As mentioned earlier, this measure is appropriate for describing any bursty source. In this section we propose another measure which can be used to judge the merits of a model, however, this measure is only appropriate for video models. To describe this measure we first need a little background on the operation of a video coder with rate control.

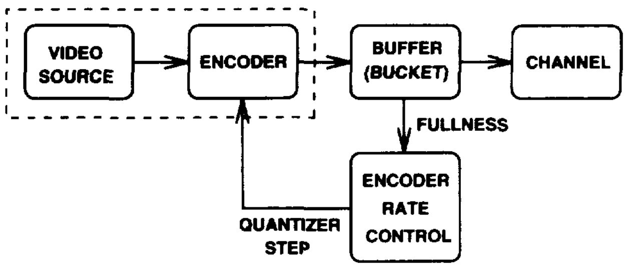


Fig. 10. Video system with rate control.

Consider a video system displayed in Fig. 10. A constant bit rate can be offered to the channel using an adaptive rate control. Bits generated by the encoder are placed in a buffer before being transmitted at a constant rate onto the channel. To ensure that the buffer neither overflows or empties, the buffer fullness is monitored in real time. If the buffer starts to fill, the rate control specifies that the encoder increase the quantization step size (q) used for the discrete cosine transform coefficients, which reduces the amount of information per frame and hence the output bit rate. As the buffer fullness decreases, the rate control decreases q which increases the output bit rate.

The above description also holds for a VBR video source if we assume that the network will monitor the traffic using a leaky bucket with specified parameters and that the source does not want any traffic identified as excessive by the network. Now, the encoder output is transmitted directly to the channel, and to ensure that the offered VBR traffic is compliant with the leaky bucket parameters, the video system again does rate control by monitoring the content of the leaky bucket and adjusting the quantization step size appropriately. Therefore, the rate control operation is identical when either a physical or a logical buffer is being monitored.

Clearly changing the quantization step size affects the quality of the video signal. If most of the step sizes used are small, then more information is retained and a higher resolution video signal is transmitted. Conversely, larger quantization implies poorer video quality. As a first step to quantifying the quality of the VBR signal, we propose to ignore the serial correlations and just look at the marginal distribution of the quantization sizes. In general, comparing two quantization histograms for different video streams may not allow us to say that one has a higher quality than the other, but if a histogram is shifted towards smaller quantization sizes then we could claim that it represents a better quality signal. Quantization histograms have been used [9] to evaluate the performance of various buffer control policies.

To simplify modeling the effect of the q-step on the bit rate output by the video encoder, we ignore many factors, including the effects of memory when changing q and the effect of changing q for different image content. We use an existing model of the variable bit rate produced by a video encoder with constant q-step (as in Section II, where $q = 8$ here), and approximate the gross effect of changing the q-step using an empirically-determined scaling factor, as shown in Fig. 11. This enables the model with rate control to be independent of both the channel rate and the specific rate control algorithm.

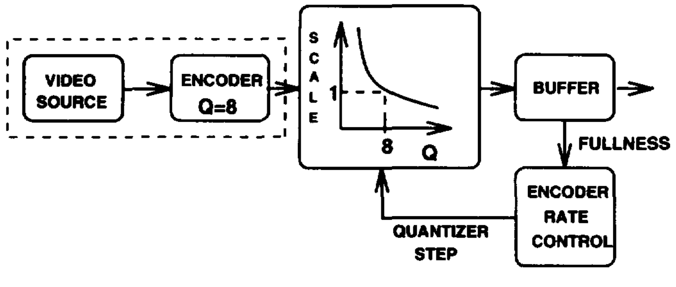


Fig. 11. Approximation to video system.

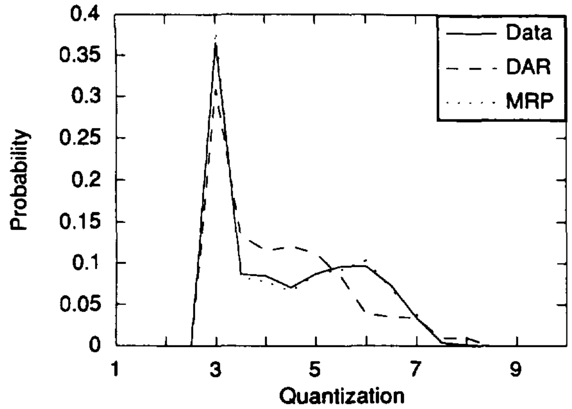


Fig. 12. Quantization histogram.

As an illustration of goodness-of-fit metric we show in Fig. 12 the quantization histogram generated by the original data and by the MRP and DAR models. Here, the nominal average rate was 768 kbps, and the buffer size is 76.8 kbits. The quantization size is chosen according to RM8, [10] with a decision made 10 times per frame. The q tabulated in the histogram is the average quantization step size used throughout the frame.

The MRP tracks the histogram almost perfectly whereas the DAR model overestimates the proportion of quantizer step sizes in the middle range and underestimates the proportion in the higher range. Thus the DAR model would give an overoptimistic estimate of the quality of the video signal. This again seems intuitive since we have seen earlier that the DAR model is less bursty than the original data.

References

[1] "Traffic control and congestion control in B-ISDN," CCITT Recommendation I.371, Geneva, Switzerland, June, 1992.

[2] W. Verbeist, L. Pinno, and B. Voeten, "The impact of the ATM concept on video coding," *IEEE J. Select. Areas Commun.*, vol. SAC-6, Dec., 1988.

[3] M. Nomura, T. Fujii, and N. Ohta, "Basic characteristics of variable bit rate video coding in ATM environment," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 752-760 June, 1989.

[4] B. Maglaris, D. Anastassiou, P. Sen, and J. D. Roberts, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, pp. 834-843, July 1988.

[5] D. Heyman, A. Tabatabai, and T.V. Lakshman, "Statistical analysis and simulation study of video teleconference traffic in ATM networks," *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 2, No. 1, March 1992.

[6] W. Whitt, "Approximating a point process by a renewal process: the view through a queue - an indirect approach," *Manage. Sci.* vol. 27, no. 6, pp. 619-36, June 1981.

[7] E. Çinlar, "Markov Renewal Theory," *Adv. Appl. Prob.* , vol. 1, pp. 123-87, 1969.

[8] A. E. Eckberg, D. M. Lucantoni, and D. T. Luan, "An approach to controlling congestion in ATM networks," *Int. J. Digit. Analog Commun. Syst.*, vol. 3, pp. 199-209, April/June 1990.

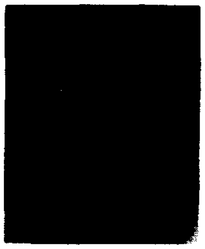
[9] J. Zdepski, D. Raychaudhuri, and K. Joseph, "Statistically based buffer control policies for constant rate transmission of compressed digitial video," *IEEE Trans. Commun.*, vol. 39, pp. 947-957, June 1991.

[10] Description of Reference Model 8, CCITT SG XV, Doc. 525, 1989.

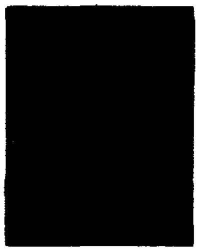
David M. Lucantoni received the B.S. degree in mathematics from Towson State University, Baltimore, MD, in 1976 and received the M.S. degree in statistics in 1978, and the Ph.D. degree in operations research in 1981, both from the University of Delaware, Newark, DE. He was awarded the Allan P. Colburn Prize for the best dissertation in the Engineering and Mathematical Sciences at the University of Delaware in 1982.

From 1981 to 1994 he was with AT&T Bell Laboratories where he was a Distinguished Member of Technical Staff. His work has included the overload control design and performance analysis of mobile telephone systems, switching systems and integrated voice and data networks. In March, 1994, he joined Motorola to work on performance aspects of the Iridium project. His current research interests are in the areas of the algorithmic analysis of stochastic models and queueing theory, and in flow and congestion controls in data communication networks.

Dr. Lucantoni was the co-recipient of the IEEE Communications Society Stephen O. Rice Prize Paper Award in the Field of Communication Theory in 1986.



Marcel F. Neuts (M'88) is Professor of Systems and Industri al Engineering at the University of Arizona, where he directs the Laboratory for Algorithmic Research, since 1985. He has previously held academic positions at Purdue University and the University of De laware. He has extensively written on stochastic models, with particular emphasis on their algorithmic analysis. His principal contribution to date is the development of extensive matrix- analytic methods for queues. In addition to over 100 journal articles, he has written a text book on probability (1973) and two research related books (1981,1989). A Book of Problems in algorithmic probability is in press. Marcel Neuts is Founding Editor of Stochastic Models and Contributing Editor of the Journal and the Advances in Applied Probability.



Amy R. Reibman (S'83–M'87) was born in Schenectady, NY on April 17, 1964. She received the B.S., M.S., and Ph.D. degrees in electrical engineering from Duke University in 1983, 1984, and 1987, respectively.

From 1988 to 1991, she was an assistant professor in the Department of Electrical Engineering at Princeton University. She is currently a Member of the Technical Staff in the Visual Communcations Research Department at AT&T Bell Laboratories. She is the Technical Program Chair for the Sixth International Workshop on Packet Video, in Portland Oregon, September 1994. Her research interests include video compression and packet video.

Dr. Reibman is a member of Sigma Xi, Eta Kappa Nu, and Tau Beta Pi.