STATISTICAL MULTIPLEXING

Kavitha Chandra
Center for Advanced
Computation and Telecommunications
University of Massachusetts at Lowell
Lowell, Massachusetts

1. INTRODUCTION

Twenty-first century communications will be dominated by intelligent high-speed information networks. Ubiquitous access to the network through wired and wireless technology, adaptable network systems, and broadband transmission capacities will enable networks to integrate and transmit media-rich services within specified standards of delivery. This vision has driven the standardization and evolution of broadband integrated services network (B-ISDN) technology since the early 1990s. The narrowband ISDN proposal for integrating voice, data, and video on the telephone line was the precursor to broadband services and asynchronous transport mode (ATM) technology. The ISDN and B-ISDN recommendations are put forth in a series of documents published by the International Telecommunication Union Telecommunication Standardization Sector (ITU-T), which was formerly CCITT [1,2] and the ATM Forum study groups [3,4]. The concept of multiplexing integrated services traffic on a common channel for efficient utilization of the transmission link capacity is central in the design of ISDN and ATM networks. ATM in particular advocates an asynchronous allocation of time slots in a time-division multiplexed frame for servicing the variable-bit-rate (VBR) traffic generated from video and data services. ATM multiplexing relies on the transport of information using fixed-size cells of 53 bytes in length and the application of fast cell-switching architectures made possible by advances in digital technology. It utilizes the concepts of both circuit and packet switching by creating virtual circuits that carry VBR streams generated by multiplexing ATM cells from voice, video, and data sources.

The ATM architecture is designed to efficiently transport traffic sources that alternate between bursts of transmission activity and periods of no activity. It also supports traffic sources with continuously changing transmission rates. One measure of traffic burstiness is the ratio of peak to average rate of the source. A circuit-switched network would conservatively allocate to each source a capacity equal to its peak rate. In this case, full resource utilization takes place only when all of the sources transmit at their peak rates. This is typically a low-probability event when the sources are statistically independent of each other. A statistical multiplexer, however, allocates a capacity that lies between the average and peak rates and buffers the traffic during periods when demand exceeds channel capacity. The process of buffering the multiplexed stream smooths the relatively high variations in the traffic rate of individual sources. The multiplexed traffic is expected to have a smaller variance about the mean rate in the limit as the number of sources multiplexed increase to a large value. As a result, there is a diminishing magnitude in the probability of occurrence of source rates that are greater than the available capacity. This feature leads to the economies of scale paradigm of statistical multiplexing that is at the core of B-ISDN and ATM transmission technologies.

Statistical multiplexers have been integral components in packet switches and routers on data networks since the 1960s. They have gained increased prominence since 1990 with the availability of broadband transmission speeds exceeding 155 Mbps and ranging upto 10 Gbps in the core of the network. In conjunction with gigabit switching speeds, the new-generation of internetworks have the hardware infrastructure for delivering broadband services. However, since broadband traffic features are highly unpredictable, the control of service quality such as packet delay and loss probabilities must be managed by a suite of intelligent and adaptable protocols. The development of these techniques is the present focus of standards bodies, researchers, and developers in industry. New Internet protocols and services are currently being proposed by the Internet Engineering Task Force (IETF) to enable integrated access and controlled delivery of multimedia services on the existing Internet packet-switching architecture [5,6]. These include integrated (Intserv) and differentiated (Diffserv) services [7], multiprotocol label switching (MPLS) [8], and resource reservation protocols (RSVPs) [9]. It is expected that ATM and the Internet will coexist with ATM infrastructure deployed at the corporate, enterprise, and private network levels. The Internet will continue to serve connectivity on the wide-area network scale. The design and performance of these new protocols and services will depend on the traffic patterns of voice, video, and data sources and their influence on queues in statistical multiplexers. These problems have been the focus of numerous studies since 1990. This article is organized as follows. Section 2 describes stochastic traffic descriptors and models that have been applied to characterize voice, video, and data traffic. In Section 3 the methods applied for performance analysis of queues driven by the aforementioned models are discussed. Section 4 concludes with a discussion on the open problems in this area.

2. TRAFFIC DESCRIPTORS AND MODELS

The characterization of network traffic with parametric models is a basic requirement for engineering communications networks. Statistical multiplexers in particular are modeled as queueing systems with finite buffer space, served by one or more transmission links of fixed or varying capacity. The service structure typically admits packets of multiple sources on a first-come first-serve (FCFS) basis. Priority-based service may also be implemented in ATM networks and more recent invocations of the Internet protocol. The statistical multiplexing gain (SMG) is an

Wiley Encyclopedia of Telecommunications, Edited by John G. Proakis
ISBN 0-471-36972-1

important performance metric that quantifies the multiplexing efficiency. The SMG may be calculated as the ratio of the number of VBR sources that can be multiplexed on a fixed capacity link under a specified delay or loss constraint and the number of sources that can be supported on the basis of peak rate allocation. To determine and maximize the SMG, admission control rules are formulated that can relate traffic characteristics to performance constraints and system parameters.

In this section, the analytic, computational, and empirical approaches for modeling traffic are discussed. A more detailed taxonomy of traffic models is presented by Frost and Melamed [10] and by Jagerman et al. [11]. The traffic is assumed to be composed of discrete units referred to as packets. The packets arriving at the multiplexer input are characterized using the sequence of random arrival times $T_1, T_2, \ldots, T_n$ measured from an origin assumed to be zero. The packets are associated with workloads $W_1, W_2, \ldots, W_n$ that may also be random variables. These workloads can represent variable Internet packet sizes fixed ATM cell sizes, or in case of batch arrivals, where more than one packet may arrive at a time instant, the workload represents the batch size. The packet interarrival times $\tau_n = T_n - T_{n-1}$ or the counting process $N(t)$, which represents the number of packets arriving in the interval $(0, t]$ are representative and equivalent descriptors of the traffic.

The most tractable traffic models result when interarrival times and workload sequences are independent random variables and independent of each other. A renewal process model is readily applicable as a traffic model in such a case. Telephone traffic on circuit-switched networks has been shown to be adequately modeled by independent negative exponential distributions for the interarrival times and call holding times. As shown by A. K. Erlang in his seminal study [12] of circuit-switched telephone traffic, the Poisson characteristics of teletraffic greatly simplify the analysis of queueing performance. Packet traffic measurement studies since 1970 have, however, shown that the arrival process of data, voice, and video applications rarely exhibit temporal independence. Traffic studies [13–15] conducted during the Arpanet days examined data traffic generated by user dialogs with distributed computer systems and showed that computer terminals transmitted information in bursts that occurred at random time intervals. Pawlita [16] presented a study of four different user applications in data networks and identified bursty traffic patterns, clustered dialog sequences and hyperexponential distributions for the user dialog times. Traffic measurement studies conducted on local-area networks [17–19] and wide-area networks [20] have found similar statistics in the packet interarrival times.

A measurement and modeling study of traffic on a token ring network by Jain and Routhier [17] showed that the packet arrivals occurred in clusters, for which they proposed a packet train model. The time between packet clusters was found to be a function of user access times, whereas the intracluster statistics were a function of the network hardware and software. More recent analyses of Internet traffic by Paxson and Floyd [21] and Caceres et al. [22] have shown that packet interarrival times generated by protocol-based applications such as file transfer, network news protocol, simple mail transfer, or remote logins are neither independent nor are they exponentially distributed.

Meier-Hellstern et al. [23], in their study of ISDN data traffic, have shown that the interarrival times for a user's terminal generated packet traffic can be modeled by superposing a gamma and power-law type probability density functions. The traffic generated in an Ethernet local area network of workstations has been shown by Gusella [24] to be nonstationary and characterized by a long-tailed interarrival time distribution. Leland et al. [25] analyzed aggregated Ethernet traffic on several timescales. A self-similar process was proposed as a model based on scale invariant features in the traffic. This model implies that the traffic variations are statistically similar over many, theoretically infinite ranges of timescales. As a result, one observes temporal dependence in the traffic structures over large time intervals. Erramilli et al. [26] propose a deterministic model based on chaotic maps for modeling these long-range dependence features. A compilation of references to work done on self-similar traffic modeling can be found in the study by Willinger et al. [27]. The aforementioned data traffic studies indicate that temporal dependence features found in measurements must be described accurately by traffic models. In this regard, static traffic descriptors such as the first- and second-order moments and marginal distributions of the traffic have been proposed. Dynamic models that capture some of the temporal features of the arrival process have also been proposed.

Traffic bursts are structures characterized by a successive occurrence of several short interarrival times followed by a relatively long interarrival time. This feature has been characterized using simple first-order descriptors such as the ratio of peak to average rate. In terms of the random interarrival times τ, the coefficient of variation c_τ captures the dispersion in the traffic through the ratio of the standard deviation and the expectation of the interarrival times

$$c_\tau = \frac{\sigma[\tau]}{E[\tau]} \tag{1}$$

Alternately, the index of dispersion $I_N(t)$ of the counting process $N(t)$ can be calculated for increasing time intervals of length t. This is a second-order characterization that captures the burstiness as a function of the variance of the process and is given by

$$I_N(t) = \frac{\text{Var }[N(t)]}{E[N(t)]} \tag{2}$$

An index of dispersion for intervals $I_\tau(n)$ may be similarly defined by replacing the numerator and denominator of Eq. (2) by the variance and expectation of the sum of n successive interarrival times. The correlation in the workloads may also be characterized using the aforementioned indices of dispersion. The magnitude and rate of increase in these traffic descriptors can capture succintly, the degree of correlation in the arrival process. An increasing magnitude of the index of dispersion with the observation time indicates highly correlated streams that are in turn linked to large packet delays and packet losses. For example, the expected number in a single server queue driven by Poisson arrivals and a general service time distribution (M/G/1) is given by the Pollaczek–Khintchine mean value formula [28], which shows that the average queue size increases in direct proportion to the square coefficient of variation of the service times. For stationary arrival processes, the limiting values of the indices of dispersion as n and t tend to infinity are shown [24] to be related to the normalized autocorrelation coefficients $\rho_\tau(j)$ $j = 0, 1, 2, \ldots$, as

$$I_N = I_\tau = {c_\tau}^2 \left[1 + 2\sum_{j=1}^{\infty} \rho_\tau(j) \right] \tag{3}$$

Sriram and Whitt [29] and others [30] apply the index of dispersion of counts (IDC) and intervals for examining the burstiness effects of superposed packet voice traffic on queues. The IDC of a single packet voice source approached a limiting value of 18 in comparison to a value of unity for a Poisson process. It has been shown [29] that under superposition, the magnitudes of the IDC of the multiplexed process approached Poisson characteristics for short time intervals. As the time interval increased the positive autocorrelations of the individual sources interact, leading to increased values of the IDC parameter. The larger the number of sources superposed, the larger is the time interval at which the superposed process deviates from Poisson-like statistics. These concepts showed the importance of identifying a relevant timescale for the superposed traffic that allows the sizing of the buffers in a queue. Although the index of dispersion descriptors have proved useful for evaluating the burstiness property in a qualitative way, they have limited application for deriving explicit measures of queue performance.

A method for estimating an index of dispersion of the queue size using the peakedness functional is presented by Eckberg [31,32]. The "peakedness" of the queue represents the ratio of the variance and expectation of the number of busy servers in an infinite server system driven by a stationary traffic process. This approach incorporates second-order traffic descriptors. The traffic workloads represented by random service times S are modeled by the service time distribution $F(t)$, its complement $Q(t) = 1 - F(t) = \Pr[S > t]$, and the autocorrelation of $Q(t)$ denoted by $R_Q(t) = \int_0^\infty Q(x)Q(t + x)\,dx$. In addition, the arrival process may be characterized by a time-varying, possibly random, arrival rate $\lambda(t)$ and its covariance density $k(\tau)$. An arrival process characterized by a random arrival rate belongs to the category of doubly stochastic processes. Cox and Lewis [33] derive the covariance density of a doubly stochastic arrival process as $k(\tau) = \sigma_\lambda^2 \rho_\lambda(\tau)$, where σ_λ^2 and $\rho_\lambda(\tau)$ are the variance and normalized autocovariance functions of $\lambda(t)$, respectively. With the traffic specified by these functions, the expected value and variance of the number of busy servers $L(t)$ at a time t may be obtained as follows:

$$E[L(t)] = \int Q(t - \tau)\lambda(\tau)\,d\tau \tag{4}$$

$$\begin{aligned}\text{Var}\,[L(t)] = \int [&Q(t - \tau)\{1 - Q(t - \tau)\}\lambda(\tau)\\ &+ k(\tau)R_Q(\tau)]\,d\tau\end{aligned} \tag{5}$$

The presence of correlations in the arrival rate for lags greater than zero cause the arrival process to be overdispersed relative to Poisson processes with constant arrival rate. The degree of dispersion is proportional to the magnitude of $k(\tau)$, $\tau > 0$ and the decay rate of $Q(t)$. This increased traffic variability has an impact on the problem of resource allocation and engineering. The peakedness functional $Z[F] = \frac{E[L(t)]}{\text{Var}\,[L(t)]}$ provides a measure of the influence of traffic variance and correlations on queue performance. $Z[F]$ has a magnitude that is greater than one for processes with nonzero $k(s)$, $s > 0$. The peakedness of the process influences the traffic engineering rules used for sizing system resources. The application of peakedness to estimate the blocking probability of finite server systems is presented by Fredericks [34]. Here a knowledge of $Z[F]$ is used in Hayward's approximation, an extension of Erlang's blocking formula to estimate the additional servers needed for $Z[F] > 1$. The utility of the peakedness characterization for analysis of delay systems has been discussed by Eckberg [32].

A more descriptive representation of the multiplexer queues is provided by the steady-state probability distribution of the buffer occupancy. If the random variable X represents the buffer occupancy in the steady state, the shape of the complementary queue distribution $G(x) = \Pr[X > x]$ provides information on the timescales at which traffic burstiness and correlations impact the queue. Livny et al. [35] have shown that the positive autocorrelations in traffic have significant impact in generating increased queue sizes and blocking probabilities relative to independent identically distributed processes. In this context it is useful to differentiate between two types of queue phenomena: queues arising from packet or cell level congestion and those arising from burst level congestion [36]. Packet level queues occur due to an instantaneous arrival of packets from different sources in the same time-slot resulting in a cumulative rate that is greater than the service capacity. It may be due to the chance occurrence of a set of interarrival times of different sources that cause individual packets to collide in time at the multiplexer input. This phenomenon can also occur for deterministic traffic, such as periodic sources, when the starting epochs are randomly displaced from one another [37,38]. Queues arising from packet-level congestion are typically of small to moderate size and can be accommodated using small buffers.

Larger queue sizes result when multiple traffic sources start transmitting in the burst state. Here, sustained transmission of a number of sources at the peak rate leads to a buildup in the queue size for time durations that are functions of the burst state statistics. Since individual times of packet arrivals are not important in this case, burst-level queues have been analyzed using the fluid flow model [39]. In the fluid approximation, the discrete packet arrival process and the buffer occupancy variables are replaced by real-valued random processes. Although burst level congestion leads to lower probability events than does packet level congestion, the decay rate of these probabilities are functions of both the service rate and traffic source burst statistics. Figure 1 shows a depiction of a typical structure of the packet- and burst-level queue components in $G(x)$. In the design of a statistical multiplexer the size of buffer is typically set to absorb packet-level queues. Burst-level queues estimated from infinite buffer queue analysis can be used to approximate the losses that take place in a finite buffer system. Finite buffer systems typically require more complex analysis than infinite buffer systems.

The fluid approximation requires that the time variation and correlation of the arrival rate process be prescribed. Finite-state Markov chain models of traffic have been applied extensively in fluid buffer analysis. Discrete- and continuous-time Markov chains (DTMC, CTMC) with finite-state space [40] are among the simplest extensions to the renewal process model for incorporating temporal dependence. The traffic correlation structure exhibits geometric or exponential rate of decay for the discrete and continuous-time Markov chains, respectively. A K-state discrete time Markov chain $Y[n], n = 0, 1, 2, \ldots$ resides in one of K states $S_1, S_2, \ldots, S_K$ at any given time n. By the Markov property, the probability of transitioning to a particular state at time n is a function of the state of the process at $n - 1$ only. These one-step transition probabilities are specified in a K-dimensional matrix P_Y as elements $p_{ij} = \Pr[Y[n] \in S_j | Y[n-1] \in S_i]$. The elements in each row of P_Y sum to unity. In a continuous time Markov chain, the transition rates are captured by an infinitesimal generating matrix Q_Y containing elements q_{ij} that represent the transition rate from state S_i to S_j for $i \neq j$. In this case, the sum of the rates in each row is equal to zero. The probability transition matrix and the generator matrix uniquely determine the rate of decay in the autocorrelations of the Markov chain. In the context of modeling traffic arrivals the transition matrix is supported by a K-state rate vector that describes the arrival rate when the traffic is in a particular state. This feature allows the variable rate features of network traffic to be represented. The rate vector in the simplest case is a set of constants that may represent the average traffic rate in each state. More general models based on a stochastic representation for the rate selection have also been considered. The Markov modulated Poisson process (MMPP) [30,41,42] is one example where the Markov process is characterized by a state-dependent Poisson process. These Markovian models of traffic can capture time variations in the arrival rate and associate these variations with a temporal correlation envelope that is determined by the magnitude of the transition probabilities. These models, however, cannot address nonexponential trends in the correlation function. To accommodate more general shapes of the correlation functions, Li and Hwang [43,44] propose the application of linear systems analysis using power spectral representation of traffic.

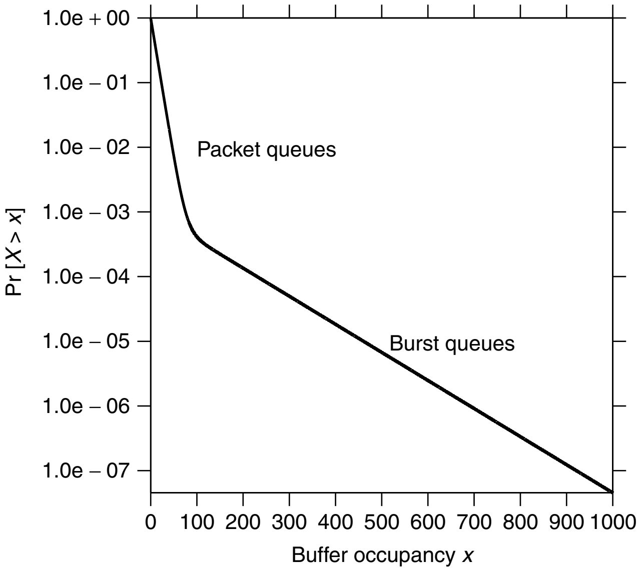

Figure 1. Packet- and burst-level components of queue size distribution.

3. PERFORMANCE OF STATISTICAL MULTIPLEXERS

Statistical multiplexing is designed to increase utilization of a resource that is subject to random usage patterns. In this work, the resource is considered to be a transmission link of finite capacity. Multiple sources access the channel on a first-come first-serve or priority basis and are allowed to queue in a buffer when the channel is busy. Statistical multiplexing gains come at the expense of a packet loss or delay probability that is considered tolerable for the applications being transported. The performance constraint is typically specified by the acceptable probability of loss for a given buffer size B. In some cases an infinite buffer queue is analyzed for tractability and the loss probability is approximated by the tail probability of queue lengths $P(X > B)$. For the limiting case of zero buffer size, the probability of loss may be calculated by determining the probability of the aggregate input rate exceeding the capacity.

The approaches to performance analysis in the literature may be classified by applications. The multiplexing of packet voice with data using two state Markov chains to model the ON and OFF states and the application of analytic and simulation-based performance analysis was the subject of numerous studies since the 1970s. With the availability of larger transmission capacities and standardization of encoding schemes for digital video, the transport and multiplexing of packet video became an active area of research in the early 1990s. Models for variable-bit-rate (VBR) video were found to be more complex and of higher dimensionality. The Markov representation for packet video typically required a large state space to capture the temporal variations and amplitude distributions. As a result, several approximation methodologies such as effective bandwidth formalisms and large deviations analyses were proposed to relate the traffic characteristics, performance constraints, and statistical multiplexer parameters. These approximations have been particularly useful in formulating admission control decisions for multiservice networks.

In the following section a review of voice/data multiplexing schemes is provided first. This is followed by a presentation of video traffic models and the analysis of their multiplexing performance using fluid buffer approximations. This leads to a discussion of admission control algorithms with particular focus on the effective bandwidth approximations.

3.1. Voice and Data Multiplexers

In the early 1980s integrated services digital networks (ISDNs) [45] were envisoned to support multiplexed transport of voice-, data-, and image-based applications on a common transport infrastructure that included both telephone and data networks. The digital telephone network with a basic transmission rate of 64 kbps (kilobits per second) was considered to be the dominant transport network. Different local user interfaces were standardized to connect end systems such as telephones, data terminals, or local area networks to a common ISDN channel. At any given time, the traffic generated on this link could be a mix of data, voice, and associated signaling and control information. The performance requirements of voice and data traffic [46] govern the design of the multiplexing system. The buffer overflow probability is a chief concern for data transmission, whereas bounding transmission delay is critical for speech signals [47]. Initial studies on the performance of voice/data multiplexing systems assumed fixed duration time-division multiplexing frames in which time slots were distributed between voice and data packets. In this context, the multiplexing efficiency of voice and data has been analyzed using various approaches that involve moving frame boundaries between voice and data slots [48–50], separate queueing buffers for voice and data [51], encoder control for voice [52], application of circuit-switching concepts for both voice and data [53], and hybrid models of circuit-switched voice/packet-switched data [54]. Maglaris and Schwartz [55] describe a variable-frame multiplexer that admits long messages of variable length and single packets that arrive as a Poisson process. The Poisson model is assumed to impose a degree of traffic burstiness on the otherwise continuous rate process. The ability to adapt frame sizes in response to traffic variations showed improved performance in terms of bandwidth utilization and delays relative to fixed-frame movable-boundary schemes. The system requirements for multiplexing data during silence periods of speech is presented by Roberge and Adoul [56]. In this work the accurate discrimination of speech and data signals is proposed using statistical pattern classification algorithms based on zero-crossing statistics of the quadrature-amplitude-modulated speech signal. A speech-data transition detector is proposed for detecting switching points in time with accuracy.

With the evolution of fast packet switching devices, the more recent approaches to voice/data integration have examined the performance of asynchronous multiplexing on a single high-speed channel. Voice/data integration concepts were motivated in part, by the traffic characteristics of data applications such as Telnet, File Transfer Protocol (FTP), and Simple Mail Transfer Protocol (SMTP). These applications generate bursts of activity separated by random durations of inactive periods. Speech patterns in telephone conversations are also characterized by random durations of talk spurts that are followed by silence periods. Brady [57] presented experimental measurements of the average durations of ON and OFF periods and transition rates between these states from a study of telephone conversations. Typically the average speech activity is found to range from 28 to 40% of the total connection time and is a function of users, language, and other such factors. Average length of talk and silence spurts are in the 0.4–1.2-s and 0.6–1.8-s ranges.

A two-state Markov process [58] representing the ON and OFF states has been the canonical model for characterizing speech-based applications, although the silence durations are seldom exponentially distributed. The characteristic transition rates between ON and OFF states can be significantly different for voice and data sources. The alternating talk spurts and silence durations of speech applications exhibit relatively slow transition rates, allowing the data to be multiplexed in the OFF periods. Data sources exhibit faster transitions between active and inactive states. A problematic feature in packet voice traffic is its temporal correlation which is induced by speech encoders and voice activity detectors [29,30]. As a result of multiplexing with voice in a queue, the departing data flow takes on the characteristics of the superposed voicestream. This feature influences the performance of other multiplexers in the transmission path.

Heffes and Lucantoni [30] model the dependence features of multiplexed voice and data traffic using a two-state Markov modulated Poisson process (MMPP). Asynchronous voice data multiplexing of MMPP sources is examined by evaluating the delay distributions of a single-server queue with first-in first-out (FIFO) service and general service time distribution. The application of this model for evaluation of overload control algorithms is discussed. Sriram and Whitt [29] extract the dependence features of aggregate voice packet arrival process from a highly variable renewal process model of a single voice source. The aggregation of multiple independent voice sources is examined using the index of dispersion of intervals (IDI). The motivation behind this approach is that the limiting value of IDI as number of sources tend to infinity completely characterizes the effect of the arrival process on the congestion characteristics of a FIFO queue in heavy traffic. This work also shows that the positive dependence in the packet arrival process is a major cause of congestion in the multiplexer queue at heavy loads. Buffer sizes larger than a critical value as determined by the characteristic correlation time scale will allow a sequence of dependent interarrival times to build up the queue, causing congestion. Limiting the size of the buffer, at the cost of increased packet loss is proposed as an approach for controlling congestion. To control packet loss that occurs from dependence in arrival process, Sriram and Lucantoni [59] propose dropping the least significant bits in the queue when the queue length reaches a given threshold. They show that under this approach the queue performance is comparable to that of a Poisson traffic source. These pioneering studies provided a comprehensive understanding on the efficiency of synchronous and asynchronous approaches for multiplexing voice and data traffic on a common channel. A quantitative characterization of the dependence features in traffic was shown to be one of the most important requirements for performance evaluation. In this regard, finite-state Markov processes were found to be amenable in both capturing some of the dependence features and allowing tractable analysis of the multiplexer queues. These studies also had limitations in that traffic measurements and measurement based models did not play a prominent role in analysis of multiplexers. However, with transition from ISDN to B-ISDN and the recognition that simple two-state Markov models are inadequate for broadband sources, more emphasis has been placed on measurement based analysis of video and data traffic. The developments in video models and multiplexers are discussed next.

3.2. Video Models and Multiplexers

Video communication services are important bandwidth consuming applications for B-ISDN. In an early study, Haskell [60] showed that multiplexing outputs of picturephone video encoders into a common buffer could achieve significant multiplexing gains. Although current compression techniques for digital video can achieve video bit rates of acceptable quality in the range of 1–5 Mbps, when hundreds of such flows are to be transported, efficient multiplexing schemes are still required. Figure 2 depicts a comparison of the temporal variation of measured video frame rates for a low-activity videoconference encoded with H.261 standard and high-activity MPEG-2 encoded entertainment video. The signals represent the number of bits in each encoded video frame as a function of the frame index. The large dispersions about a mean rate are evident, as are the sudden transitions in frame rate amplitudes when encoding changes from predictive to refresh mode.

The transport of variable bit rate (VBR) video using statistical multiplexing has been examined in numerous studies. VBR video is preferred over traditional constant-bit-rate (CBR) video due to the improved image quality and shorter delays at the encoder. Statistical multiplexing invariably results in buffering delays and losses, which can significantly degrade video quality. To minimize the amount of delay and loss, the networking community has focused on the development of effective and implementable congestion control schemes, including connection admission control and usage parameter control. To minimize the impact of delay and loss, the video community has focused on developing good error concealment algorithms and designing efficient two-layer coding algorithms [61,62] for use in combination with the dual-priority transport provided by ATM networks. For example, while one-layer MPEG-2 produces generally unacceptable video quality with a cell loss ratio of 10^{-3}, losses at this rate with SNR scalability (one of the four standardized layered coding algorithms of MPEG-2) are generally invisible, even to experienced viewers [63].

Various application- and coding-specific models of one-layer VBR video have been proposed in the literature. Maglaris et al. [64] was among the first to analyze short (10-s) segments of low-activity videophone signals. A first order autoregressive (AR) process was proposed for the number of bits in successive video frames of a single source. The multiplexed video was modeled by a birth–death Markov process. In this model, transitions are limited to neighboring states. The model parameters were selected to match the mean and short-term covariance structure in the measurements. The states of the Markov chain

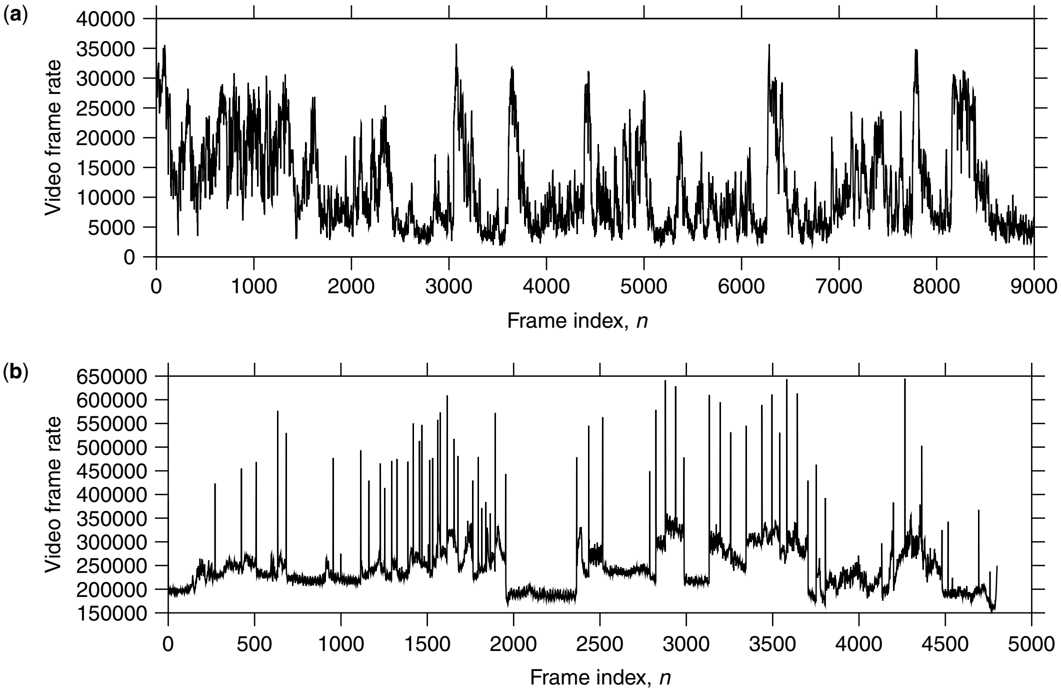

Figure 2. Sample paths of VBR video in videoconferencing and entertainment applications.

were derived by quantizing the aggregate source rate histogram into a fixed number of levels. A choice of 20 levels per source was found adequate for the low-activity videoconferencing source. Sen et al. [65] extended this model to accommodate moderate activity sources, using additional states to model low and high activity levels. The resulting source model is equivalent to that obtained by superposition of independent ON–OFF processes. Grunenfelder et al. [66] used a conditional replenishment encoder that exhibits strong correlations effects. The superposed video process was assumed to be wide sense stationary. The multiplexer was modeled using a general arrival process with independent arrivals and deterministic service times. A source model for full motion video was presented by Yegenoglu et al. [67] using an autoregressive process with time-varying coefficients. The selection of the coefficients was based on the state of a discrete-time Markov chain. The transition and rate matrices were constructed by matching the rate probability density function with that obtained from measurements. Moderate-activity videoconference data were also modeled by Heyman et al. [68]. The rate evolution was modeled by a Markov chain with identical transition probabilities in each row. These probabilities were modeled as a negative binomial distribution. The number of states in the model was of the order of the peak rate scaled by a factor of 10. This ranged from 400 to 500 states. The within state correlations were modeled by a discrete AR (DAR) process that resulted in a diagonally dominant matrix structure. This structure did not model single-source statistics very well since there were no selective transitions between states based on source characteristics. The DAR model failed to capture the short-term correlations in the traffic of a single source. Lucantoni et al. [69] proposed a Markov renewal process (MRP) model for VBR source traffic. Results of this model show that burstiness in video data can be captured more accurately than in the DAR model. Although the MRP was shown to perform better than the DAR model in capturing the burstiness, the MRP still did not match the cell loss probabilities for large buffer sizes. Heyman and Lakshman [70] studied high-activity video sources and concluded that their DAR model proposed for videoconference sources could not be applied as a general model to all sources. Skelly et al. [71] also used Markov chains to verify a histogram-based queueing model for multiplexing. They determined, on the basis of simulation, that a fixed number of eight states were sufficient to model the video source.

The video encoding system effects play a significant role in shaping the temporal and amplitude variation of compressed video. Traffic shaping algorithms at policing systems that enforce constraints on the output rate of the encoding system also play an important role in shaping digital video traffic [72–74]. In general, the coder that produces a bit stream conforming to constraints will not have the same statistical characteristics of an unconstrained coder. The idea that encoders could be constrained to generate traffic described by Markov chains was explored by Heeke [75] for designing better traffic policing and control algorithms. Pancha and Zarki [76] have examined the traffic characteristics resulting from various combinations of the quantization parameter, the inter-to-intraframe ratio and the priority breakpoint in MPEG one-layer and two-layer encoding, respectively. Data generated for each parameter set were modeled by a Markov chain by selecting the number of states based on the ratio of peak rate to the standard deviation of the frame rates. Frater et al. [77] verify the performance of a non-Markovian model for full motion video based on scene characterization by matching the cell loss probabilities at different buffer sizes. Krunz and Hughes [78] modeled MPEG, with distinct models for each different frame type. The selection of an adequate number of states in the Markov chain model of video such as to adequately model the spectral content is discussed by Chandra and Reibman [79]. Adequate spectral content in single sources is found necessary to understand the scaling aspects of Markov models under multiplexing.

Non-Markovian models exhibiting long-range dependence have been proposed by Garrett and Willinger [80] and others [77,81]. Ryu and Elwalid [82] show that long-term correlations do not significantly affect network performance over a reasonable range of cell losses, buffer sizes, and network operating parameters. Grossglauser and Bolot [83] also propose that correlation timescales to be considered in the traffic depend on the operating parameters and that full long-range dependence characterization of traffic is unnecessary. The impact of temporal correlation in the output rate of a VBR video source on the queue response has been examined [43], and it has been shown that macrolevel correlations can be modeled by Markov-chain-based models. Long-range dependence seen in VBR video has also been examined and the queueing results have been compared to those obtained using the DAR model [84]. It was concluded that for moderate buffer sizes, the short-range correlations obtained using Markov chain models are sufficient to estimate the buffer characteristics.

The aforementioned studies have determined that correlations on many timescales are an inherent feature in video sources and that Markov modulated source models are appropriate for capturing these dynamics. The ubiquitous use of multistate Markov models has led to work on their performance in queues. The application of finite-state Markov chain video traffic models for H.261 and MPEG coders in simulation studies [79,85] has shown that with an increase in the number of multiplexed video sources and corresponding increase in the channel capacity, the loss probability can be significantly reduced and reasonable multiplexing gains achieved. It was shown that typically 15–20 states are required to faithfully model the queue behavior of moderate to high-activity video sources. When using too few states, the tail probabilities of the rate histogram will not be captured, thereby yielding an underestimate of the packet delay or loss probability. This situation has been observed by Hasslinger [86] in modeling VBR sources using semi-Markov models.

The performance analysis of queues driven by large-dimensional Markovian traffic sources may be approached using exact queueing analysis in discrete time and discrete state space. This approach becomes quickly intractable as the number of sources increase due to exponential increase in state dimension. In the limit of a large number of sources operating in the heavy-traffic regime, the discrete arrival and departure times may be replaced by a fluid approximation. This analysis technique is discussed in the next section.

3.3. Fluid Buffer Models

Fluid flow models assume that the packet arrival process at a multiplexer occurs continuously in time and may be characterized by continuous random fluctuations in the arrival rate [87]. This approach is applicable when the packet sizes are small relative to the link capacity. The computational model presented by Anick et al. [39] affords the estimation of the delay and loss distributions in multiplexers fed by Markov modulated fluid sources and served at constant rate. In this method, the buffer occupancy X is assumed to be a continuous valued random variable. The arrival process of each source is represented by a finite-state continuous-time Markov generator Q and associated diagonal rate matrix R. If K is the number of states required to represent a single source, and N is the number of sources (assumed identical) being multiplexed, the superposition can be modeled by the Markov generator Q_N and diagonal rate matrix R_N, which are computed as the N-fold Krönecker sums $Q \oplus Q \oplus \cdots \oplus Q$ and $R \oplus R \cdots \oplus R$, respectively. The Krönecker sum operation increases the dimension of the multiplexed source generator matrix to $M = K^N$.

The aggregated traffic stream enters a queue with finite or infinite waiting room. Packets in the buffer are serviced on a first-in first-out basis at a constant service rate. The cumulative probability distribution of the buffer occupancy x in steady state is specified by the row vector $\vec{p}$: $[p_0(x), p_1(x), \ldots, p_{M-1}(x)]$, where element $p_i(x) = \text{Prob}\ [X \leq x; \textit{source in state } i]$. For a service rate of C packets per second, the probabilities satisfy the equation

$$\frac{\partial \vec{p}}{\partial x} D = \vec{p} Q_N \tag{6}$$

where the matrix $D = [R_N - IC]$ captures the drift from the service rate in each state. Here I is the identity matrix. The solution of Eq. (6) follows that of an eigenvalue problem and may be represented in terms of the tail probability distribution $G(x)$ as

$$G(x) = \Pr[X > x] = \sum_{i=0}^{M-1} a_i(x) e^{-z_i x} \tag{7}$$

where $z_i,\ i = 0, \ldots, M-1$ are the eigenvalues of the matrix $Q_N D^{-1}$. The coefficients $a_i(x)$ are functions of the eigenvalues and eigenvectors [88,89] of $Q_N D^{-1}$. For an infinite buffer, subject to consideration that the solution is bounded at x equal to infinity, the coefficients of the exponentially growing modes are set equal to zero. The amplitudes of the remaining modes are determined by applying the appropriate boundary conditions for overload and underload states. Underload states represented by states of the drift matrix with negative elements are subject to the condition $p_i(x = 0) = 0$. The coefficients for overload states are solved by equating $p_i(\infty)$ to the steady-state probability of the multiplexed source being in state i. For a buffer of finite size, all of the eigenvalues are retained and boundary condition at infinity replaced by the corresponding value at the buffer size.

The aforementioned approach requires the solution of an eigenvalue problem for a matrix whose upper bound on dimension scales as $O(K^N)$. To counter this dimensionality problem, reduced order traffic models have been used as approximations. For two-state ON–OFF Markov processes the superposition yields a generator of $O(N)$ states. Multi-state Markov sources have been approximated by the superposition of multiple two-state ON–OFF sources by matching first and second moments of the two processes [64,65,90]. The number of two-state sources selected for this model is often an arbitrarily choice. For moderate to high-activity video sources, this approximation can be shown to underestimate the packet delays.

Correlation effects afforded by the generator matrix play a dominant role in structuring the features of burst-level queueing delays. A traffic source represented by a finite-state Markov chain exhibits temporal autocorrelations that decay exponentially in time. The rate of decay is governed by the dominant eigenvalues of the generator matrix Q. It can be shown that the characteristic correlation time-scale of a single source is retained in the superposed traffic. The selection of an adequate single source model order K that captures all of the dominant modes is therefore an important consideration in building the traffic model. High-activity video sources often require K to be in the range of 15–20 states. Figure 3 shows the effect of choosing an inadequate number of states K for modeling the H.261 encoded videoconference source shown in Fig. 2a. As K is increased from 5 to 16, the asymptotic decay rate of the complementary delay distribution approaches that exhibited by the measurements.

For sources with large-dimensional K, even for moderate values of N, the estimation of buffer occupancy distributions for the multiplexed system becomes computationally intensive. A method for reducing the state-space dimension of multiplexed source generator is given by Thompson et al. [91]. The reduction process involved the quantization of the rates and the fundamental rate was chosen to yield the best match to the mean and variance of the rate. States having equal rates were aggregated, thereby reducing the number of states in the generator matrix. The resulting model allowed for scalable analysis as the number of sources was increased.

For large-dimensional systems, asymptotic approximations to the model given in Eq. (7) may be obtained for large buffer sizes and small delay or loss probabilities [92]. In this approximation

$$G(x) \sim \alpha e^{-\beta x} \quad x \to \infty \tag{8}$$

where α is referred to as the *asymptotic constant* and β is the largest negative eigenvalue of the matrix $Q_N D^{-1}$. The asymptotic decay rate $-\beta$ is a function of the service rate C and may be determined with relative ease. The asymptotic constant, however, requires knowledge

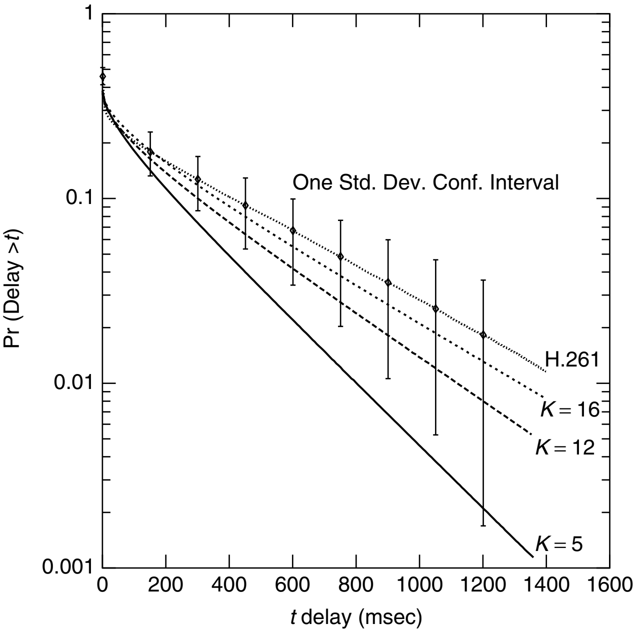

Figure 3. Influence of number of states K chosen to model video source.

of all the eigenvectors and eigenvalues of the system. Approximate methods for estimating α for Markovian systems have been derived [93,94]. Although most of the asymptotic representations have considered Markovian sources, there have been some results for traffic modeled as stationary Gaussian processes [95] and fractional Brownian motion [96].

Very often, $\alpha = G(0)$ is assumed to be unity in the heavy-traffic regime. This allows a very usable descriptor of multiplexer performance that is referred to as the *effective bandwidth* of a source, which assumes the limiting form of the tail probabilities be structured as

$$G(x) \approx e^{-\beta x} \tag{9}$$

To achieve a specified value of β that satisfies given performance constraints, the required capacity may be shown [97] to be obtained as the maximal eigenvalue of the matrix $[R_N + \frac{Q_N}{\beta}]$. This capacity is referred to as the effective bandwidth (EB) of the multiplexed source. In the limit as β approaches zero and ∞, EB approaches the source average and peak rate, respectively. However, as noted by Choudhury et al. [98], the EB approximation can lead to conservative estimates for bursty traffic sources that undergo significant smoothing under multiplexing. It was shown that the asymptotic constant was itself asymptotically exponential in the number of multiplexed sources N. For traffic sources with indices of dispersion greater than Poisson, this parameter decreased exponentially in N, reflecting the multiplexing gain of the system. The application of these approximations in designing efficient multiplexing systems through admission control is discussed next.

3.4. Effective Bandwidths and Admission Control

Network traffic measurements have identified application and system dependent features that influence the traffic characteristics. Characterization in terms of the mean traffic rate is inadequate because of the large variability in its value over time. As traffic mixes and their performance requirements change, network mechanisms that adapt to these variations are of critical importance in broadband networks. Admission and usage control policies are two proposals in place in ATM networks and the next-generation Internet An admission control process determines if a source requesting connection can be admitted into the system without perturbing the performance of existing connections. This control algorithm should therefore make its decision taking into account a specific set of traffic parameters, available link capacity, and performance specifications. If a flow is admitted, the usage control policy monitors the flow characteristics to ensure that its bandwidth usage is within the admitted values.

To facilitate admission control, the concept of service classes has been introduced to categorize traffic with disparate traffic and service characteristics. Multiplexing among the same and across different service classes have been analyzed. The performance of five classes of admission control algorithms is reviewed by Knightly and Shroff [99]. These classes include scheduling based on average and peak rate information [100], effective bandwidth calculations [97,101,102], their refinements from large deviation principles [103], and maximum variance approaches based on estimating the upper tails of Gaussian process models of traffic [95]. A overview of the EB and its refinements is presented, since it appears to be the most generally applicable formalism.

It is assumed that K classes of traffic are to be admitted into a node served by a link capacity C packets per second. If N_i sources of type i exist, each characterized by an effective bandwidth E_i, then the simplest admission policy is given by the linear control law:

$$\sum_{i=1}^{K} N_i E_i \leq C \tag{10}$$

The effective bandwidths are derived taking into consideration the traffic characteristics and performance requirements of each class and available capacity C. Defining the traffic generated by type i source on a timescale t by a random variable $A_i[0,t]$, the effective bandwidth derived from large-deviation principles is given by [103]

$$E_i(s,t) = \frac{1}{st} \log E[e^{sA_i[0,t]}] \tag{11}$$

where the parameter s is related to the decay rate of $G(x)$ and captures the multiplexing efficiency of the system. It is calculated from the specified probability of loss or delay bounds [89]. The term on the right is the log moment generating function of the arrival process. The workload can be described over a time t that represents the typical time taken for the buffer to overflow starting from an empty state. For a fixed value of t, the EB is an increasing function of s and lies between the mean and peak values of $A_i[0, t]$. This may be shown by a Taylor series approximation of E_i as $s \to 0$ and $s \to \infty$ respectively. Methods for deriving E_i for different traffic classes are discussed by Chang [104]. The aforementioned model assumes that all the multiplexed sources have the same quality of service requirements. If not, all sources achieve the performance of the most stringent source. Kulkarni et al. [105] consider an extension of this approach for addressing traffic of multiple classes. For the superposition, since the total workload is given by $A[0, t] = \sum_{i=1}^{K} A_i[0, t]$, the effective capacity C_e of the multiplexed system is

$$C_e = \sum_{i=1}^{K} E_i \tag{12}$$

The admission control algorithm simply compares C_e with available capacity C and if $C_e < C$ allows the new source to be admitted into the system.

4. CONCLUSIONS AND OPEN PROBLEMS

The current status on statistical multiplexing in broadband telecommunication networks has been presented. The multiplexing issues in the early 1980s were concerned with voice and data integration on 63-kbps telephone channels. Variations of synchronous time-division multiplexing using moving boundaries between voice and data slots, silence detection for insertion of data packets, and development of adaptive speech encoders were the primary concerns in designing efficient multiplexers. The transition to broadband era characterized by capacities exceeding 155 Mbps evolved with the design and standardization of asynchronous transfer mode networks. ATM networks were envisioned to integrate and optimize features of circuit- and packet-switched networks. The increase in switching speeds and network capacities in the 1990s and the invention of the World Wide Web concept, accelerated the development of many new applications and services that involved networked voice, video, and data. With increased accessibility of the Internet, many of the ATM related paradigms such as traffic characterization, admission control, and statistical multiplexing efficiency are now more relevant for the public Internet.

The important open issues at present are robust characterization of traffic and the derivation of traffic models that can be tractably analyzed in a queueing system. With the automation of end systems and application of complex encoders and detectors, the traffic patterns seen on networks today may not readily map to a pure stochastic model framework. On the contrary, traffic measurements indicate that deterministic patterns and nonlinearities in traffic amplitudes are new prevailing features in broadband networks. Large-dimensional stochastic models are seen to be required to capture these features. The computational complexity associated in analyzing the multiplexing problem for such models has led to some innovative approximation techniques. The characterization of a traffic source by an effective bandwidth is one such result that is derived by application of large-deviation principles. Derived in the asymptotic limit of large number of multiplexed sources, large buffer sizes and link capacities and small probabilities of delay and loss, effective bandwidths offer a conservative, but computationally feasible model for evaluating multiplexing efficiency in theory. The design of real-time algorithms for applying these concepts on a network and discovery of their effectiveness is expected to be the next step in the statistical multiplexing analysis.

BIOGRAPHY

Kavitha Chandra received her B.S. degree in electrical engineering in 1985 from Bangalore University, India, her M.S. and D.Eng. degrees in computer and electrical engineering from the University of Massachusetts at Lowell in 1987 and 1992, respectively. She joined AT&T Bell Laboratories in 1994 as a member of technical staff in the Teletraffic and Performance Analysis Department. From 1996 to 1998 she was a senior member of technical staff in the Network Design and Performance Analysis department of AT&T Laboratories. She is currently an associate professor in the Department of Electrical and Computer Engineering and principal faculty in the Center for Advanced Computation and Telecommunications at the University of Massachusetts at Lowell. Dr. Chandra received the Eta Kappa Nu Outstanding Electrical Engineer Award (honorable mention) in 1996 and the National Science Foundation Career Award in 1998. Her research interests are in the areas of network traffic and performance analysis, wireless networks, acoustic and electromagnetic wave propagation, adaptive estimation and control.

BIBLIOGRAPHY

1. CCITT Red Book, *Integrated Services Digital Network (ISDN), Series I Recommendation*, Vol. III, Fascicle III.5, 1985.
2. ITU-T, *Recommendation i.113. Vocabulary of Terms for Broadband Aspects of ISDN*, Vol. Rev. 1, Geneva, 1991.
3. ATM Forum, *User-Network Interface (UNI) Specification Version 3.1*, 1994.
4. ATM Forum, *Traffic Management Specification Version 4.0*, 1996.
5. D. D. Clark, S. Shenker, and L. Zhang, Supporting real-time applications in an integrated services packet network: Architecture and mechanism, *Proc. SIGCOMM 92*, 1992, pp. 14–26.
6. T. Chen, Evolution to the programmable Internet, *IEEE Commun. Mag.* **38**: 124–128 (2000).
7. G. Eichler, Implementing integrated and differentiated services for the Internet with ATM networks: A practical approach, *IEEE Commun. Mag.* **38**: 132–141 (2000).
8. D. Awduche, MPLS and traffic engineering in IP networks, *IEEE Commun. Mag.* **37**: 42–47 (1999).
9. L. Zhang et al., RSVP: A new resource reservation protocol, *IEEE Network* **7**(5): 8–18 (1993).

10. V. S. Frost and B. Melamed, Traffic modeling for telecommunications networks, *IEEE Commun. Mag.* **32**(4): 70–81 (1994).

11. D. Jagerman, B. Melamed, and W. Willinger, Stochastic modeling of traffic processes, in J. Dshalalow, ed., *Frontiers in Queuing: Models, Methods and Problems*, CRC Press, 1996.

12. A. K. Erlang, The theory of probabilities and telephone conversations, *Nyt Tidsskrift Matematik B* **20**: 33 (1909). (English translation in E. Brockmeyer, H. L. Halstrom and A. Jensen (1948), The life and works of A. K. Erlang, The Copenhagen Telephone Company, Copenhagen.

13. E. G. Coffman and R. C. Wood, Interarrival statistics for time sharing systems, *Commun. ACM* **9**: 5000–5003 (1966).

14. P. E. Jackson and C. D. Stubbs, A study of multi-access computer communications, *AFIPS Conf. Proc.* **34**: 491–504 (1969).

15. E. Fuchs and P. E. Jackson, Estimates of distributions of random variables for certain computer communications traffic models, *Commun. ACM* **13**: 752–757 (1970).

16. P. F. Pawlita, Traffic measurements in data networks, recent measurement results, and some implications, *IEEE Trans. Commun.* **29**(4): 525–535 (1981).

17. R. Jain and S. A. Routhier, Packet trains- measurements and a new model for computer network traffic, *IEEE J. Select. Areas Commun.* **4**(6): 986–995 (1986).

18. J. F. Schoch and J. A. Hupp, Performance of the Ethernet local network, *Commun. ACM* **23**(12): 711–721 (1980).

19. D. N. Murray and P. H. Enslow, An experimental study of the performance of a local area network, *IEEE Commun. Mag.* **22**: 48–53 (1984).

20. F. A. Tobagi, Modeling and measurement techniques in packet communication networks, *Proc. IEEE* **66**: 1423–1447 (1978).

21. V. Paxson and S. Floyd, Wide-area traffic: The failure of Poisson modeling, *IEEE/ACM Trans. Network.* **3**(3): 226–244 (1996).

22. R. Caceres, P. Danzig, S. Jamin, and D. Mitzel, Characteristics of wide-area Tcp/Ip conversations, *Proc. ACM/SIGCOMM*, 1991, pp. 101–112.

23. K. S. Meier-Hellstern, P. E. Wirth, Y. Yan, and D. A. Hoeflin, Traffic models for ISDN data users: Office automation application, in A. Jensen and V. B. Iversen, eds., *Teletraffic and Data Traffic, a Period of Change*, Elsevier Science Publishers, 1991, pp. 167–172.

24. R. Gusella, Characterizing the variability of arrival processes with indexes of dispersion, *IEEE J. Select. Areas Commun.* **9**(2): 203–211 (1991).

25. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Trans. Network.* **2**(1): 1–15 (1994).

26. A. Erramilli, R. P. Singh, and P. Pruthi, Chaotic maps as models of packet traffic, *Proc. 14th Int. Teletraffic Congress*, 1994, Vol. 1, pp. 329–338.

27. W. Willinger, M. S. Taqqu, and A. Erramilli, A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks, in F. P. Kelly, S. Zachary, and I. Ziedins, eds., *Stochastic Networks: Theory and Applications*, Oxford Univ. Press, 1996, pp. 339–366.

28. R. B. Cooper, *Introduction to Queueing Theory*, 3rd ed., CEEPress Books, 1990.

29. K. Sriram and W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data, *IEEE J. Select. Areas Commun.* **4**(6): 833–846 (1986).

30. H. Heffes and D. M. Lucantoni, A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE J. Select. Areas Commun.* **4**: 856–868 (1986).

31. A. E. Eckberg, Jr., Generalized peakedness of teletraffic processes, *10th Int. Teletraffic Congress*, 1983.

32. A. E. Eckberg, Jr., Approximations for bursty (and smoothed) arrival queueing delays based on generalized peakedness, in *11th Int. Teletraffic Congress*, 1985.

33. D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Chapman & Hall, 1966.

34. A. A. Fredericks, Congestion in blocking systems — a simple approximation technique, *Bell Syst. Tech. J.* **59**: 805–827 (1980).

35. M. Livny, B. Melamed, and A. K. Tsiolis, The impact of autocorrelation on queueing systems, *Manage. Sci.* **39**(3): 322–339 (1993).

36. W. Roberts, ed., *Performance Evaluation and Design of Multiservice Networks*, COST 224 Final Report, Commission of the European Communities, 1992.

37. I. Norros, J. W. Roberts, A. Simonian, and J. T. Virtamo, The superposition of variable bit rate sources in an ATM multiplexer, *IEEE J. Select. Areas Commun.* **9**(3): 378–387 (1991).

38. J. W. Roberts and J. T. Virtamo, The superposition of periodic cell arrival streams in an ATM multiplexer, *IEEE Trans. Commun.* **39**: 298–303 (1991).

39. D. Anick, D. Mitra, and M. M. Sondhi, Stochastic theory of a data-handling system with multiple sources, *Bell Syst. Tech. J.* **8**: 1871–1894 (1982).

40. E. Cinlar, *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

41. H. Heffes, A class of data traffic processes-covariance function characterization and related queueing results, *Bell Syst. Tech. J.* **59**: 897–929 (1980).

42. W. Fischer and K. Meier-Hellstern, The Markov-modulated Poisson process (MMPP) cookbook, *Perform. Eval.* **18**: 149–171 (1992).

43. S. Q. Li and C. L. Hwang, Queue response to input correlation functions: Discrete spectral analysis, *IEEE/ACM Trans. Network.* **1**(5): 522–533 (1993).

44. S. Q. Li and C. L. Hwang, Queue response to input correlation functions: Continuous spectral analysis, *IEEE/ACM Trans. Network.* **1**(6): 678–692 (1993).

45. M. Decina, W. S. Gifford, R. Potter, and A. A. Robrock (guest eds.), Special issue on Integrated Services Digital Network: Recommendations and Field Trials — I, *IEEE J. Select. Areas Commun.* **4**(3): (1986).

46. J. G. Gruber and N. H. Le, Performance requirements for integrated voice/data networks, *IEEE J. Select. Areas Commun.* **1**(6): 981–1005 (1983).

47. J. G. Gruber, Delay related issues in integrated voice and data networks, *IEEE Trans. Commun.* **29**(6): 786–800 (1980).

48. N. Janakiraman, B. Pagurek, and J. E. Neilson, Performance analysis of an integrated switch with fixed or variable frame rate and movable voice/data boundary, *IEEE Trans. Commun.* **32**: 34–39 (1984).

49. A. G. Konheim and R. L. Pickholtz, Analysis of integrated voice/data multiplexing, *IEEE Trans. Commun.* **32**(2): 140–147 (1984).

50. K. Sriram, P. K. Varshney, and J. G. Shantikumar, Discrete-time analysis of integrated voice-data multiplexers with and without speech activity detectors, *IEEE J. Select. Areas Commun.* **1**(6): 1124–1132 (1983).

51. H. H. Lee and C. K. Un, Performance analysis of statistical voice/data multiplexing systems with voice storage, *IEEE Trans. Commun.* **33**(8): 809–819 (1985).

52. T. Bially, B. Gold, and S. Seneff, A technique for adaptive voice flow control in integrated packet networks, *IEEE Trans. Commun.* **28**: 325–333 (1980).

53. E. A. Harrington, Voice/data integration using circuit switched networks, *IEEE Trans. Commun.* **28**: 781–793 (1980).

54. C. J. Weinstein, M. K. Malpass, and M. J. Fisher, Data traffic performance of an integrated circuit and packet-switched multiplex structure, *IEEE Trans. Commun.* **28**: 873–877 (1980).

55. B. Maglaris and M. Schwartz, Performance evaluation of a variable frame multiplexer for integrated switched networks, *IEEE Trans. Commun.* **29**(6): 801–807 (1981).

56. C. Roberge and J. Adoul, Fast on-line speech/voiceband-data discrimination for statistical multiplexing of data with telephone conversations, *IEEE Trans. Commun.* **34**(8): 744–751 (1986).

57. P. T. Brady, A statistical analysis of on-off patterns in 16 conversations, *Bell Syst. Tech. J.* **47**: 73–91 (1968).

58. P. T. Brady, A model for generating on-off speech patterns in two-way conversations, *Bell Syst. Tech. J.* **48**: 2445–2472 (1969).

59. K. Sriram and D. M. Lucantoni, Traffic smoothing effects of bit dropping in a packet voice multiplexer, *IEEE Trans. Commun.* **37**(7): 703–712 (1989).

60. B. G. Haskell, Buffer and channel sharing by several interframe picturephone coders, *Bell Syst. Tech. J.* **51**(1): 261–289 (1972).

61. M. Ghanbari, Two-layer coding of video signals for VBR networks, *IEEE J. Select. Areas Commun.* **7**: 771–781 (1989).

62. S. Tubaro, A two layers video coding scheme for ATM networks, *Signal Process. Image Commun.* **3**: 129–141 (1991).

63. R. Aravind, M. R. Civanlar, and A. R. Reibman, Packet loss resilience of mpeg-2 scalable coding algorithms, *IEEE Trans. Circuits Syst. Video Technol.* **6**: 426–435 (1996).

64. B. Maglaris et al., Performance models of statistical multiplexing in packet video communications, *IEEE Trans. Commun.* **36**(7): 834–844 (1988).

65. P. Sen, B. Maglaris, N. Rikli, and D. Anastassiou, Models for packet switching of variable bit-rate video sources, *IEEE J. Select. Areas Commun.* **7**(5): 865–869 (1989).

66. R. Grunenfelder, J. P. Cosmos, S. Manthorpe, and A. Odinma-Okafor, Characterization of video codecs as autoregressive moving average processes and related queueing system performance, *IEEE J. Select. Areas Commun.* **9**: 284–293 (1991).

67. F. Yegenoglu, B. Jabbari, and Y. Zhang, Motion classified autoregressive modeling of variable bit rate video, *IEEE Trans. Circuits Syst. Video Technol.* **3**: 42–53 (1993).

68. D. Heyman, A. Tabatbai, and T. V. Lakshman, Statistical analysis and simulation study of video teletraffic in atm networks, *IEEE Trans. Circuits Syst. Video Technol.* **2**: 49–59 (1992).

69. D. M. Lucantoni, M. F. Neuts, and A. R. Reibman, Methods for performance evaluation of VBR video traffic models, *IEEE/ACM Trans. Network.* **2**: 176–180 (1994).

70. D. P. Heyman and T. V. Lakshman, Source models for VBR broadcast-video traffic, *IEEE/ACM Trans. Network.* **4**: 40–48 (1996).

71. P. Skelly, M. Schwartz, and S. Dixit, A histogram based model for video traffic behavior in an ATM multiplexer, *IEEE/ACM Trans. Network.* **1**(4): 447–459 (1993).

72. P. Pancha and M. El Zarki, Prioritized transmission of VBR MPEG video, *Proc. GLOBECOM'92*, 1992, pp. 1135–1139.

73. M. R. Ismail, I. E. Lambadaris, M. Devetsikiotis, and A. R. Raye, Modelling prioritized MPEG video using tes and a frame spreading strategy for transmission in ATM networks, *Proc. INFOCOM'9*, 1995, Vol. 5, pp. 762–769.

74. A. R. Reibman and A. W. Berger, Traffic descriptors for VBR video teleconferencing, *IEEE/ACM Trans. Network.* **3**: 329–339 (1995).

75. H. Heeke, A traffic control algorithm for ATM networks, *IEEE Trans. Circuits Syst. Video Technol.* **3**(3): 183–189 (1993).

76. P. Pancha and M. El. Zarki, Bandwidth allocation schemes for variable bit rate MPEG sources in ATM networks, *IEEE Trans. Circuits Syst. Video Technol.* **3**(3): 192–198 (1993).

77. M. R. Frater, J. F. Arnold, and P. Tan, A new statistical model for traffic generated by VBR coders for television on the broadband isdn, *IEEE Trans. Circuits Syst. Video Technol.* **4**(6): 521–526 (1994).

78. M. Krunz and H. Hughes, A traffic model for MPEG-coded VBR streams, *Perform. Eval. Rev.* (*Proc. ACM SIGMETRICS'95*) **23**: 47–55 (1995).

79. K. Chandra and A. R. Reibman, Modeling one- and two-layer variable bit rate video, *IEEE/ACM Trans. Networks* **7**(3): 398–413 (1999).

80. M. W. Garrett and W. Willinger, Analysis, modeling and generation of self-similar VBR video traffic, *Proc. ACM SIGCOMM'94*, 1994, pp. 269–280.

81. J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, Long range dependence in variable bit-rate video traffic, *IEEE Trans. Commun.* **43**: 1566–1579 (1995).

82. B. K. Ryu and A. Elwalid, The importance of long-range dependence of VBR video traffic in atm traffic engineering: Myths and realities, *Proc. ACM SIGCOMM'96*, 1996, pp. 3–14.

83. M. Grossglauser and J. D. Bolot, On the relevance of long-range dependence in network traffic, *Proc. ACM SIGCOMM'96*, 1996, pp. 15–24.

84. D. Heyman and T. V. Lakshman, What are the implications of long-range dependence for VBR-video traffic engineering? *IEEE/ACM Trans. Networking* **4**: 301–317 (1996).

85. K. Chandra and A. R. Reibman, Modeling two-layer SNR scalable MPEG-2 video traffic, *Proc. 7th Int. Workshop Packet Video*, 1996, pp. 7–12.

86. G. Hasslinger, Semi-markovian modelling and performance analysis of variable rate traffic in ATM networks, *Telecommun. Syst.* **7**: 281–298 (1997).

87. L. Kleinrock, *Queueing Systems*, Vol. 2, Wiley, New York, 1976.

88. J. Walrand and P. Varaiya, *High-Performance Communication Networks*, Morgan Kaufmann, 1996.

89. M. Schwartz, *Broadband Integrated Networks*, Prentice-Hall, 1996.

90. J. W. Mark and S.-Q. Li, Traffic characterization for integrated services networks, *IEEE Trans. Commun.* **38**: 1231–1242 (1990).

91. C. Thompson, K. Chandra, S. Mulpur, and J. Davis, Packet delay in multiplexed video streams, *Telecommun. Syst.* **16**: 335–345 (2001).

92. J. Abate, G. L. Choudhury, and W. Whitt, Asymptotics for steady-state tail probabilities in structured Markov queueing models, *Stochastic Models* **10**: 99–143 (1994).

93. R. G. Addie and M. Zukerman, An approximation for performance evaluation of stationary single server queues, *IEEE Trans. Commun.* **42**: 3150–3160 (1994).

94. A. Elwalid et al., Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing, *IEEE J. Select. Areas Commun.* **13**: 1004–1016 (1995).

95. J. Choe and N. B. Shroff, A central-limit-theorem-based approach for analyzing queue behavior in high-speed networks, *IEEE/ACM Trans. Network.* **6**(5): 659–671 (1998).

96. I. Norros, On the use of Fractal Brownian motion in the theory of connectionless networks, *IEEE J. Select. Areas Commun.* **13**: 953–962 (1995).

97. A. I. Elwalid and D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, *IEEE/ACM Trans. Network.* **1**(3): 329–343 (1993).

98. G. Choudhury, D. M. Lucantoni, and W. Whitt, Squeezing the most of ATM, *IEEE Trans. Commun.* **44**(2): 203–217 (1996).

99. E. W. Knightly and N. B. Schroff, Admission control for statistical qos: Theory and practice, *IEEE Network* **13**(2): 20–29 (1999).

100. D. Ferrari and D. Verma, A scheme for real-time channel establishment in wide-area networks, *IEEE J. Select. Areas Commun.* **8**: 368–379 (1990).

101. G. Kesidis, J. Walrand, and C. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, *IEEE/ACM Trans. Network.* **1**(4): 424–428 (1993).

102. C. S. Chang and J. A. Thomas, Effective bandwidth in high-speed digital networks, *IEEE J. Select. Areas Commun.* **13**: 1019–1114 (1995).

103. F. Kelly, *Stochastic Networks: Theory and Applications*, Oxford Univ. Press, 1996.

104. C. S. Chang, Stability, queue length and delay of deterministic and stochastic queueing networks, *IEEE Trans. Automatic Control* **39**: 913–931 (1994).

105. V. G. Kulkarni, L. Gun, and P. F. Chimento, Effective bandwidth vectors for multiclass traffic multiplexed in a partitioned buffer, *IEEE J. Select. Areas Commun.* **6**(13): 1039–1047 (1995).