STOCHASTIC MODELING OF TRAFFIC PROCESSES

David L. Jagerman
NEC USA, Inc.
4 Independence Way
Princeton, NJ 08540

Benjamin Melamed
Rutgers University – RUTCOR
P.O. Box 5062
New Brunswick, NJ 08903

Walter Willinger
Bellcore
445 South Street
Morristown, NJ 07690

ABSTRACT

Modern telecommunications networks are being designed to accomodate a heterogenous mix of traffic classes ranging from traditional telephone calls to video and data services. Thus, traffic models are of crucial importance to the engineering and performance analysis of telecommunications system, notably congestion and overload controls and capacity estimation.

This chapter surveys teletraffic models, addressing both theoretical and computational aspects. It first surveys the main classes of teletraffic models commonly used in teletraffic modeling, and then proceeds to survey traffic methods for computing statistics relevant to the engineering a teletraffic network.

1 INTRODUCTION

Traffic is the driving force of telecommunications systems, representing customers making phone calls, transferring data files and other electronic information, or more recently, transmitting compressed video frames to a display device. The most common modeling context is queueing; traffic is offered to a queue or a network of queues, and various performance measures are calculated or estimated. These include queue length, server utilization, waiting times and traffic loss. Performance studies utilize traffic models in two basic ways: either as part of an analytical model, or to drive a discrete-event Monte Carlo simulation. Either way, traffic constitutes the common grist to the transport mechanism of the telecommunications system under study.

The material in this chapter is organized in two broad groupings, Section 2 and Section 3. Section 2 surveys the main model classes of traffic streams commonly used in telecommunications. The emphasis here is on the theoretical aspects of traffic models, whether used in analytical context or Monte Carlo simulation context. Section 3 surveys the main teletraffic methods, designed originally to dimension and provision primarily telephone networks, subject to a prescribed grade of service. The emphasis here is on the computational aspects of the methods surveyed.

The material in each grouping is organized by and large in ascending generality and complexity. The order of presentation also roughly tracks the corresponding chronological order of emergence of the models and methods surveyed.

2 MODELS OF TRAFFIC STREAMS

Simple traffic consists of single arrivals of discrete entities (packets, cells, etc.) It can be mathematically described as a *point process* [13, 10, 87], consisting of a sequence of arrival instants $T_1, T_2, \ldots, T_n, \ldots$ measured from the origin 0; by convention, $T_0 = 0$. There are two additional equivalent descriptions of point processes: *counting processes* and *interarrival time* processes. A counting process $\{N(t)\}_{t=0}^{\infty}$ is a continuous-time, non-negative integer-valued stochastic process, where $N(t) = \max\{n : \; T_n \leq t\}$ is the number of (traffic) arrivals in the interval $(0, t]$. An interarrival time process is a real-valued random sequence $\{A_n\}_{n=1}^{\infty}$, where $A_n = T_n - T_{n-1}$ is the length of the time interval separating the n-th arrival from the previous one. The equivalence of these descriptions follows from the fact that $T_n = \sum_{k=1}^{n} A_k$, and from the equality of events

$$\{N(t) = n\} \;=\; \{T_n \leq t < T_{n+1}\} \;=\; \{\sum_{k=1}^{n} A_k \leq t < \sum_{k=1}^{n+1} A_k\}. \qquad (1)$$

The interarrival times, $\{A_n\}$, are assumed to form a stationary sequence, unless otherwise stated. An alternative characterization of point processes, called *stochastic intensity* theory [20, 46], is briefly discussed in Section 2.7.

Compound traffic consists of batch arrivals; that is, arrivals may consist of more than one unit at an arrival instant T_n. In order to fully describe compound traffic, one also needs to specify a real-valued random sequence $\{B_n\}_{n=1}^{\infty}$, where B_n is the (random) number of units in the batch. At a higher level of abstraction, B_n may represent some general attributes of the n-th arrival, e.g., the amount of "work" associated with the n-th arrival or its itinerary in a network. Such compound traffic processes, called *marked point processes* [31], are outside the scope of this paper.

Discrete-time traffic processes correspond to the case when time is slotted. Mathematically, this means that the random variables A_n can assume only integer values, or equivalently, that the random variables $N(t)$ are allowed to increase only at integer T_n.

In addition to arrival times and batch sizes, it is often useful (and sometimes essential) to incorporate the notion of *workload* into the traffic description. The workload is a general concept describing the amount of work W_n brought to a system by the n-th arriving unit; it is usually assumed independent of interarrival times and batch sizes. A typical example is the sequence of service time requirements of arrivals at a queuing system, though in queueing, one usually refers to the arrival process alone as traffic. On the other hand, traffic reduces to workload description, when interarrival times are deterministic. A case in point is compressed video, also known as coded video. Video information is rarely transmitted over a network in its raw form. Rather, engineers take advantage of the considerable visual redundancy inherent in digitized pictures to compress each frame into a fraction of its original size. The compressed frames have random sizes (bit rates) which are then transported over the network and decoded at their destination. The term VBR (variable bit rate) video is used to refer to this kind of video traffic. Coded video frames (arrivals) must be delivered deterministically every 1/30 of a second or so, for high-quality video. The workload consists of coded frame sizes (say, in bits), since a frame size is roughly proportional to its transmission time (service requirement).

The following notation will be used. The common distribution function of the A_n is denoted by $F_A(x)$. Similarly, $\lambda_A = 1/E[A_n]$ denotes the traffic rate, $\sigma_A^2 = Var[A_n]$, and

$c_A = \lambda_A \sigma_A$. Unless otherwise stated, it is assumed that $0 < \sigma_A < \infty$, and that $\{A_n\}$ is simple, namely, $P\{A_n = 0\} = 0$. A traffic stream is denoted by X when a particular traffic description (via A, N or T) is immaterial. In that case, traffic parameters may also be subscripted by X, e.g., λ_X is equivalent notation for λ_A, and similarly for other traffic parameters.

2.1 RENEWAL TRAFFIC MODELS

This section briefly touches on renewal traffic processes and the important special cases of Poisson processes and Bernoulli processes.

Renewal models have a long history, due to their relative mathematical simplicity. In a renewal traffic process the A_n are iid (independent, identically distributed), but their distribution is allowed to be general. Unfortunately, with few exceptions, the superposition of independent renewal processes does not yield a renewal process. Those which do, occupy a special position in traffic theory and practice. Historically, many queueing models routinely assumed a renewal offered traffic.

Renewal processes, while simple analytically, have a severe modeling drawback — the autocorrelation function of $\{A_n\}$ vanishes identically for all non-zero lags. The importance of capturing autocorrelations stems from the role of the autocorrelation function as a statistical proxy for temporal dependence in time series. Moreover, positive autocorrelations in $\{A_n\}$ can explain, to a large extent, the phenomenon of traffic burstiness. Bursty traffic is expected to dominate broadband networks, and when offered to a queueing system, it gives rise to much worse performance measures (such as mean waiting times) as compared to renewal traffic (which lacks temporal dependence); see [66] for a detailed discussion. Consequently, models which capture the autocorrelated nature of traffic are essential for predicting the performance of emerging broadband networks.

2.1.1 Poisson Processes

Poisson models are the oldest traffic models, dating back to the advent of telephony and the renowned pioneering telephone engineer A.K. Erlang. A Poisson process [13] can be characterized as a renewal process whose interarrival times $\{A_n\}$ are exponentially distributed with rate parameter λ, that is, $P\{A_n \leq t\} = 1 - \exp(-\lambda t)$. Equivalently, it is a counting process, satisfying $P\{N(t) = n\} = \exp(-\lambda t)(\lambda t)^n/n!$, and the number of arrivals in disjoint intervals is statistically independent (a property known as *independent increments*).

Poisson processes enjoy some elegant analytical properties. First, the superposition of independent Poisson processes results in a new Poisson process whose rate is the sum of the component rates. Second, the independent increment property renders Poisson a memoryless process. This, in turn, greatly simplifies queueing problems involving Poisson arrivals. And third, Poisson processes are fairly common in traffic applications which physically comprise a large number of independent traffic streams, each of which may be quite general. The theoretical basis for this phenomenon is known as Palm's Theorem [68], Vol. II, p. 582. It roughly states that under suitable but mild regularity conditions, such multiplexed streams approach a Poisson process, as the number of streams grows, but the individual rates decrease so as to keep the aggregate rate constant. Thus, traffic on main communications arteries are commonly believed to follow a Poisson process, as opposed to traffic on upstream tributaries, which are less likely to be Poisson. However, traffic aggregation (multiplexing) need not always result in a Poisson stream; see the discussion in Section 2.6 for a counter-example.

Time dependent Poisson processes are defined by letting the rate parameter λ depend on time. Compound Poisson processes are defined in the obvious way, by specifying the distribution of the batch size B_n, independent of the A_n.

2.1.2 Bernoulli Processes

Bernoulli processes are the discrete-time analog of Poisson processes (time dependent and compound Bernoulli processes are defined in the natural way). Here the probability of an arrival in any time slot is p, independent of any other one. It follows that for slot k, the corresponding number of arrivals is Binomial, i.e., $P\{N_k = n\} = \binom{k}{n} p^n (1-p)^{k-n}$. The time between arrivals is geometric with parameter p, i.e., $P\{A_n = j\} = p(1-p)^j$.

2.1.3 Phase-Type Renewal Processes

An important special case of renewal models occurs when the interarrival times are of the so-called *phase type*. Phase-type interarrival times can be modeled as the time to absorption in a continuous-time Markov process $C = \{C(t)\}_{t=0}^{\infty}$ with state space $\{0, 1, \ldots, m\}$; here, state 0 is absorbing, all other states are transient, and absorption is guaranteed in a finite time. To determine A_n, start the process C with some initial distribution π. When absorption occurs (i.e., the process enters state 0), stop the process. The elapsed time is A_n, implying that it is a probabilistic mixture of sums of exponentials. Then, restart C with the same initial distribution π, and repeat the procedure independently to get A_{n+1}.

Phase-type renewal processes give rise to relatively tractable traffic models. They also enjoy the property that they are dense in the space of all distributions of non-negative random variables, that is, any interarrival distribution can be approximated arbitrarily closely by phase-type ones; see, e.g., [83, 3].

2.2 MARKOV-BASED TRAFFIC MODELS

Unlike renewal traffic models, Markov and Markov-renewal traffic models [13] introduce dependence into the random sequence $\{A_n\}$. Consequently, they can potentially capture traffic burstiness, due to non-zero autocorrelations in $\{A_n\}$.

Consider a Markov process $M = \{M(t)\}_{t=0}^{\infty}$ with a discrete state space. In this case, M behaves as follows: it stays in a state i for an exponentially distributed holding time with parameter λ_i which depends on i alone [13]; it then jumps to state j with probability p_{ij}, such that the matrix $P = [p_{ij}]$ is a probability matrix. In a simple Markov traffic model, each jump of the Markov process is interpreted as signaling an arrival, so interarrival times are exponential, their rate parameter depending on the state from with the jump occurred.

Markov models in slotted time can be defined for the process $\{A_n\}$ in terms of a Markov transition matrix $P = [p_{ij}]$ [13]. Here, state i corresponds to i idle slots separating successive arrivals, and p_{ij} is the probablity of a j-slot separation, given that the previous one was an i-slot separation. Arrivals may be single, a batch of units or a continuous quantity. Batches may themselves be described by a Markov chain, whereas continuous-state, discrete-time

Markov processes can model the (random) workload arriving synchronously at the system. In all cases, the Markovian structure introduces dependence into interarrival separation, batch sizes and successive workloads, respectively.

Markov-renewal models are more general than discrete-state Markov processes, yet retain a measure of simplicity and analytical tractability. A Markov renewal process $R = \{(M_n, \tau_n)\}_{n=0}^{\infty}$ is defined by a Markov chain $\{M_n\}$ and its associated jump times $\{\tau_n\}$, subject to the following constraint: The distribution of the pair (M_{n+1}, τ_{n+1}), of next state and inter-jump time, depends only on the current state M_n, but not on previous states nor on previous inter-jump times. Again, if we interpret jumps (transitions) of $\{M_n\}$ as signaling arrivals, we would have dependence in the arrival process. Also, unlike the Markov process case, the interarrival times can be arbitrarily distributed, and these distributions depend on both states straddling each interarrival interval [13].

MAP (Markovian Arrival Process) is a broad and versatile subclass of Markov renewal traffic processes, enjoying analytical tractability [67]. Here, the interarrival times are phase-type (see above) but with a wrinkle: Traffic arrivals still occur at absorption instants of the auxiliary Markov process C, but the latter is not restarted with the same initial distribution; rather, the restart state depends on the previous transient state from which absorption had just occurred. While MAP is analytically simple, it enjoys considerable versatility; its formulation includes Poisson processes, phase-type renewal processes and others as special cases. It also has the appealing properties that it is dense in the set of all point processes [4], and the superposition of independent MAP traffic streams results in a MAP traffic stream governed by a Markov process whose state space is the cross product of the component state spaces [67].

2.2.1 Markov-Modulated Processes

Markov-modulated models constitute an extremely important class of traffic models. The idea is to introduce an explicit notion of state into the description of a traffic stream — an auxiliary Markov process is evolving in time and its current state controls (modulates) the probability law of the traffic mechanism.

Let $M = \{M(t)\}_{t=0}^{\infty}$ be a continuous-time Markov process, with state space of $1, 2, \ldots m$ (more complicated state spaces are possible). Now assume that while M is in state k, the probability law of traffic arrivals is completely determined by k, and this holds for every $1 \leq k \leq m$. Note that when M undergoes a transition to, say, state j, then a new probability law for arrivals takes effect for the duration of state j, and so on. Thus, the probability law for arrivals is modulated by the state of M (such systems are also called doubly stochastic, but the term "Markov modulation" makes it clearer that the traffic is stochastically subordinated to M).

Certainly, the modulating process can be more complicated than a Markov process (so the holding times need not be restricted to exponential random variables), but such models are far less analytically tractable. For example, Markov Renewal processes constitute a natural generalization of Markov-modulated processes with generally-distributed interarrival times, but those will not be reviewed here.

2.2.2 Markov-Modulated Poisson Processes

The most commonly used Markov-modulated model is the MMPP (Markov-Modulated Poisson Process) model, which combines the simplicity of the modulating (Markov) process with that of the modulated (Poisson) process. In this case, the modulation mechanism simply stipulates that in state k of M, arrivals occur according to a Poisson process at rate λ_k. As the state changes, so does the rate.

MMPP models can be used in a number of ways. Consider first a single traffic source with variable rate. A simple traffic model would quantize the rate into a finite number of rates and each rate would give rise to a state in some Markov modulating process. Certainly, it remains to verify that exponential holding times of rates are an appropriate description, but the Markov transition matrix $Q = [Q_{kj}]$ of the putative M can be easily estimated from empirical data: Simply quantize the empirical data, and then estimate Q_{kj} by calculating the fraction of times that M switched from state k to state j.

As a simple example, consider a two-state MMPP model, where one state is an "on" state with an associated positive Poisson rate, and the other is an "off" state with associated rate zero (such models are also known as interrupted Poisson for obvious reasons). These models have been widely used to model voice traffic sources [42]; the "on" state corresponds to a talk spurt (when the speaker emits sound), and the "off" state corresponds to a silence (when the speaker pauses for a break). This basic MMPP model can be extended to aggregations of independent traffic sources, each of which is an MMPP, modulated by an individual Markov process M_i, as described above. Let $J(t) = (J_1(t), J_2(t), \ldots, J_r(t))$, where $J_i(t)$ is the number of active sources of traffic type i, and let $M(t) = (M_1(t), M_2(t), \ldots, M_r(t))$ be the corresponding vector-valued Markov process taking values on all r-dimensional vectors with non-negative integer components. The arrival rate of class i traffic in state $(j_1, j_2, \ldots, j_r)$ of $M(t)$ is $j_i \lambda_i$.

2.2.3 Transition-Modulated Processes

Transition-modulated processes are a variation on the state modulation idea. Essentially, the modulating agent is a state transition rather than a state per se. However, note that a state transition can be described simply by a pair of states, whose components are the one before transition and the one after it.

A transition-modulated traffic model in discrete time is described in [91]; its generalization to continuous time is straightforward. Let $M = \{M_n\}_{n=1}^{\infty}$ be a discrete-time Markov process on the positive integers. State transitions occur on slot boundaries, and are governed by an $m \times m$ Markov transition matrix $P = [P_{ij}]$. Let B_n denote the number of arrivals in slot n, and assume that the probabilities $P\{B_n = k | M_n = i, M_{n+1} = j\} = t_{ij}(k)$, are independent of any past state information (the parameters $t_{ij}(k)$ are assumed given). Notice that these probabilities are conditioned on transitions (M_n, M_{n+1}) of M from state M_n to state M_{n+1} during slot n. Furthermore, the number of traffic arrivals during slot n is completely determined by the transition of the modulating chain (through the parameters $t_{ij}(k)$).

Markov-modulated traffic models are a special case of Markovian transition-modulated ones: Simply take the special case where the conditioning event is $\{M_n = i\}$. That is, $t_{ij}(k) = t_i(k)$ depends only on the state i of the modulating chain in slot n, but is independent of its state j in the next slot $n+1$. Conversely, Markovian transition-modulated processes can be thought of as Markov-modulated ones, but on a larger state space. Indeed, if $\{M_n\}$ is Markov, so is the process $\{(M_n, M_{n+1})\}$ of its transitions.

As before, multiple transition-modulated traffic models can be defined, one for each traffic class of interest. The complete traffic model is obtained as the superposition of the individual traffic models. For queueing studies in discrete time, another wrinkle is the assignment of priorities to different classes, so as to order their arrivals in a buffer [91].

2.3 FLUID TRAFFIC MODELS

The fluid traffic paradigm dispenses with individual traffic units. Instead, it views traffic as a stream of fluid, characterized by a flow rate (e.g., bits per second), so that a traffic count is replaced by a traffic volume.

Fluid models are appropriate to cases where individual units are numerous relative to a chosen time scale. Put differently, an individual unit is by itself of vanishingly little significance, just as one molecule more or less in a water pipeline has but an infinitesimal effect on the flow. In the B-ISDN (Broadband Integrated Services Digital Networks) context of ATM (Asynchronous Transfer Mode), all packets are fixed size cells of relatively short length (53 bytes); in addition, the high transmission speeds (say, on the order of gigabit/second) render the transmission impact of individual cells negligible. The analogy of a cell to a fluid molecule is a plausible one. To further highlight this analogy, contrast an ATM cell with a much bigger transmission unit, say, a compressed high-quality video frame, which consists on the order of a thousand cells. A traffic arrival stream of coded frames should be modeled as a discrete stream of arrivals, since such frames are typically transmitted at the rate of 30 frames per second. However, a fluid model is appropriate for the constituent cells.

An important advantage of fluid models is their conceptual simplicity. But important benefits will also accrue to a simulation model of fluid traffic. To see that, consider again a broadband ATM scenario. If one is to distinguish among cells, then each of them would have to count as an event. The time granularity of event processing would be quite fine, and consequently, processing cell arrivals would consume vast CPU and possibly memory resources, even on simulated time scales of minutes. A statistically meaningful simulation may often be infeasible. A fluid simulation would assume that the incoming fluid flow remains (roughly) constant over much longer time periods. Traffic fluctuations are modeled by events signaling a change of flow rate. As these changes can be assumed to happen far less frequently than individual cell arrivals, one can realize enormous savings in computing. In fact, infeasible simulations of cell arrival models can be replaced by feasible simulations of fluid models of comparable accuracy. In a queueing context, it is easy to manipulate fluid buffers. Furthermore, the waiting time concept simply becomes the time it takes to serve (clear) the current buffer. Since fluid models assume a deterministic service rate, these statistics can be readily computed. Typically, though, larger traffic units (say coded frames) are of greater interest than individual cells. Modeling the larger units as discrete traffic and their transport as fluid flow will give us the best of both worlds: we can measure waiting times and enjoy significant savings on simulation computing resources.

Typical fluid models [2, 55] assume that sources are bursty — of the "on-off" type. While in the "off" state, traffic is switched off, whereas in the "on" state traffic arrives deterministically at a constant rate λ. For analytical tractability, the durations of "on" and "off" periods are assumed to be exponentially distributed and mutually independent (that is, they form an alternating renewal process). A Markov model of a set of quantized (fluid) traffic rates is presented in [84]. Fluid traffic models of these types can be analyzed as Markov-modulated constant rate traffic. The host of generalizations, described above for MMPP, carries over to fluid models as well, including multiple sources and multiple classes of sources.

2.4 AUTOREGRESSIVE-TYPE TRAFFIC MODELS

Autoregressive-type models define the next variate in the sequence as an explicit function of previous variates (from the same times series or a related one) within a time window stretching from the present into the past. Such models are particularly suitable for modeling VBR coded video — a projected major consumer of bandwidth in emerging high-speed communications networks. The nature of video frames is such that successive frames within a video scene vary visually very little (recall that there are 30 frames per second in a high-quality video). Only scene changes (and other visual discontinuities) can cause abrupt changes in frame bit rate. Thus, the sequence of bit rates (frame sizes) comprising a video scene may be modeled by an autoregressive scheme, while scene changes can be modeled by some modulating mechanism, such as a Markov chain. However, see Sections 2.5 and 2.6 for alternative modeling approaches.

2.4.1 Linear Autoregressive (AR) Processes

The class AR(p) consists of linear *autoregressive* models of order p,

$$X_n \;=\; a_0 \;+\; \sum_{r=1}^{p} a_r\, X_{n-r} \;+\; \epsilon_n, \quad n > 0, \tag{2}$$

where $(X_{-p+1}, \ldots, X_0)$ is a prescribed random vector (usually a multivariate normal vector), the a_r, $0 \le r \le p$, are real constants, and the ϵ_n are zero-mean, uncorrelated random variables (white noise), called *residuals*, which are independent of the X_n [5]. In a good model, the residuals ought to be of smaller magnitude than the X_n, in order to "explain" the empirical data.

The recursive form of (2) makes it clear how to generate the next random element in the sequence $\{X_n\}_{n=0}^{\infty}$ from previous ones. This simplicity makes them popular candidates for modeling autocorrelated traffic. A simple AR(2) model was used in [43] to model VBR coded video. More elaborate models can be constructed out of AR(p) models combined with other schemes. For example, in [80], the video bit rate traffic was modeled as a sum $R_n = X_n + Y_n + K_n\, C_n$, where the first two terms comprise independent AR(1) schemes, and the third term is a product of a simple Markov chain and an independent Normal variate from an iid Normal sequence. The purpose of having two autoregressive schemes is to achieve a better fit of the empirical autocorrelation function; the third term is designed to capture sample path spikes due to video scene changes.

2.4.2 Moving Average (MA) Processes

The class MA(q) consists of *moving average* models of order q,

$$X_n \;=\; \sum_{r=0}^{q} b_r\, \epsilon_{n-r}, \quad n > 0, \tag{3}$$

where the b_r, $0 \le r \le q$, are real constants, and the ϵ_n are zero-mean uncorrelated random variables [5]. MA models are autocorrelated time series, since successive variates are defined in terms of common subsets of ϵ_n.

2.4.3 Autoregressive Moving Average (ARMA) Processes

Modeling stationary and invertible processes using AR or MA processes often calls for the estimation of a large number of parameters, thereby reducing estimation efficiency. To mitigate this problem, one may try to combine (2) and (3) in a mixed model of the form

$$X_n \;=\; a_0 \;+\; \sum_{r=1}^{p} a_r\, X_{n-r} \;+\; \sum_{r=0}^{q} b_r\, \epsilon_{n-r}. \tag{4}$$

The class of models defined by (4) is denoted by ARMA(p, q), and referred to as *autoregressive moving average* models of order (p, q). Clearly, ARMA modeling is more flexible than AR modeling or MA modeling alone.

2.4.4 Autoregressive Integrated Moving Average (ARIMA) Processes

Related to ARMA(p, q), is the class ARIMA(p, d, q) of *autoregressive integrated moving average processes*, obtained by replacing X_n in (4) by the d-th differences of the process $\{X_n\}$. ARIMA modeling is more general than ARMA modeling; its scope includes certain types of non-stationary series. The term "integrated" alludes to the fact that an ARMA model is fitted to the differenced data, and that ARMA model is then "integrated" (summed) to yield the target (non-stationary) ARIMA model. For example, an ARMA(p, q) process may be viewed as an ARIMA($p, 0, q$) process and the random walk model can be viewed as an ARIMA($0, 1, 0$) process.

From a modeling vantage point, ARIMA processes constitute a broad and flexible class of stochastic models whose parameter estimation, forecasting, model identification and model selection are well-understood (e.g., see [9, 94]). Two points should be kept in mind, though, when considering ARMA and ARIMA processes as models of modern traffic. First, both ARMA and ARIMA processes have autocorrelation functions which decay geometrically in the lag, namely, $\rho(n) \sim r^n$ for some $0 < r < 1$, as $n \to \infty$. In this sense, ARMA and ARIMA processes are inherently short-range dependent models, incapable of parsimoniously capturing the persistence phenomena observed empirically in many modern high-speed networks (see Section 2.6). Secondly, it has also been observed that the corresponding empirical marginal distributions are typically not Gaussian, but rather, they tend to be skewed to the right [43, 77]. Since the theoretical relationship between skewness/kurtosis of an ARIMA process and the corresponding parameters of its generating white noise process is not yet well-understood, the effect of the marginal mismatch is not clear.

2.5 TES TRAFFIC MODELS

The need to capture traffic burstiness has motivated the TES (*Transform-Expand-Sample*) modeling approach which strives to simultaneously model both the marginal distribution and autocorrelation function of an empirical record [50, 51, 76], traffic being a special case [77]. The empirical TES methodology assumes that some stationary empirical time series (e.g., traffic measurements over time) are available. It aims to construct a model satisfying the following three fidelity requirements, simultaneously:

1. The model's marginal distribution should match its empirical counterpart (a histogram, in practice).
2. The model's leading autocorrelations should approximate their empirical counterparts up to a reasonable lag.
3. the sample paths (histories) generated by simulating the model should "resemble" the empirical time series.

The first two are precise quantitative requirements, whereas the third requirement is a heuristic qualitative one. Nevertheless, the latter is worth adopting, since sample path "resemblance" is informally used by modelers to increase the confidence in the model.

2.5.1 TES Processes

TES processes constitute a versatile family of stochastic sequences, consisting of two broad classes, called TES$^+$ and TES$^-$. The superscript (plus or minus) is a mnemonic reminder of the fact that the TES family gives rise to processes with positive and negative lag-1 autocorrelations, respectively. TES models consist of two stochastic processes in lockstep, called *background* and *foreground* sequences. Background TES sequences have the form

$$U_n^+ = \begin{cases} U_0, & n = 0 \\ \langle U_{n-1}^+ + V_n \rangle, & n > 0 \end{cases} \qquad U_n^- = \begin{cases} U_n^+, & n \text{ even} \\ 1 - U_n^+, & n \text{ odd} \end{cases} \tag{5}$$

Here, U_0 is distributed uniformly on $[0, 1)$; $\{V_n\}_{n=1}^{\infty}$ is a sequence of iid random variables, independent of U_0, called the *innovation sequence*; and angular brackets denote the modulo-1 (fractional part) operator $\langle x \rangle = x - \max\{\text{integer } n : \ n \le x\}$. Background sequences play an auxiliary role. The real target are foreground sequences of the form

$$X_n^+ = D(U_n^+), \quad X_n^- = D(U_n^-), \tag{6}$$

where D is a transformation from $[0, 1)$ to the reals, called a *distortion*.

It can be shown that all background sequences are Markovian and stationary; however, it is worth noting that while the transition structure of $\{U_n^+\}$ is stationary, that of $\{U_n^-\}$ is not (it depends on the even or odd time index). More importantly, the marginal distribution of both $\{U_n^+\}$ and $\{U_n^-\}$ is uniform on $[0, 1)$, *regardless* of the probability law of the innovations $\{V_n\}$ [50]. The *inversion method* [21] allows us to transform any background uniform variates to foreground ones with an arbitrary marginal distribution. To illustrate this idea, suppose one has an empirical time series $\{Y_n\}_{n=0}^{N}$, from which one computes an empirical density $\hat{h}_Y$ and its associated distribution function $\hat{H}_Y$. Then, the random variable $X = \hat{H}_Y^{-1}(U)$ has density $\hat{h}_Y$. Thus, TES foreground sequences can match any empirical distribution.

2.5.2 The Empirical TES Modeling Methodology

The empirical TES methodology actually employs a composite two-stage distortion of the form

$$D_{Y,\xi}(x) \;=\; \hat{H}_Y^{-1}(S_\xi(x)), \quad x \in [0,1), \tag{7}$$

where $\hat{H}_Y^{-1}$ is the inverse of the empirical histogram distribution, and S_ξ is a "smoothing" operation, called a *stitching transformation*, parameterized by $0 < \xi < 1$, and given by

$$S_\xi(y) \;=\; \begin{cases} y/\xi, & 0 \le y < \xi \\ (1-y)/(1-\xi), & \xi \le y < 1 \end{cases} \tag{8}$$

For $0 < \xi < 1$, the effect of S_ξ is to render the sample paths of background TES sequences more "continuous-looking". Because stitching transformations preserve uniformity, the inversion method via $\hat{H}_Y^{-1}$ guarantees that the corresponding foreground sequence would have the prescribed marginal distribution $\hat{H}_Y$. The empirical TES modeling takes advantage of this fact which effectively decouples the fitting requirements of the empirical distribution and the empirical autocorrelation function. Since the former is automatically guaranteed by TES theory, one can concentrate on fitting the latter. In practice, fitting is carried out by a heuristic search for pairs, (ξ, f_V), where ξ is a stitching parameter and f_V is an innovation density; the search is declared a success on finding that the corresponding TES sequence gives rise to an autocorrelation function that adequately approximates its empirical counterpart, and whose simulated sample paths bear "adequate resemblance" to their empirical counterparts.

In practice, efficient searches of this kind must rely on software support. TEStool is a visual interactive software environment designed to support TES modeling [34]. TEStool allows the user to read in empirical sample paths and calculate their empirical statistics (histogram, autocorrelation function and spectral density) in textual and graphical forms. It further provides services to generate and modify TES models and to superimpose the corresponding TES statistics on their empirical counterparts. The search proceeds in an interactive style, guided by visual feedback: each model modification triggers a recalculation and redisplay of the results. TES model autocorrelations and spectral densities are calculated numerically from fast and accurate formulas developed in [50, 51]. This activity is further simplified by restricting the innovation densities f_V to be step functions. Simple densities like that can be readily modified graphically, since steps are visually represented by rectangles, and those can be created, deleted, stretched and moved easily with the mouse.

Recently, an algorithmic modeling approach has been devised and implemented for TES modeling. The algorithm first carries out a brute-force computation over a subspace of step-function innovation densities and various stitching parameters; recall that the distortion is completely determined by the empirical record and user-supplied histogram parameters. Of those, the algorithm selects the best n combinations of pairs, (f_V, ξ), in the sense that the resulting TES model autocorrelation functions minimize the mean square error with respect to the empirical autocorrelation function. The analyst then selects among those n candidate models the one whose Monte Carlo sample paths bear the "most resemblance" to the empirical record. Experience shows that the TES modeling algorithm produces better and faster results than its heuristic counterpart.

Figure 1 depicts the application of the aforementioned TES modeling algorithm. It displays the final TEStool screen at the end of a TEStool modeling session performed on an

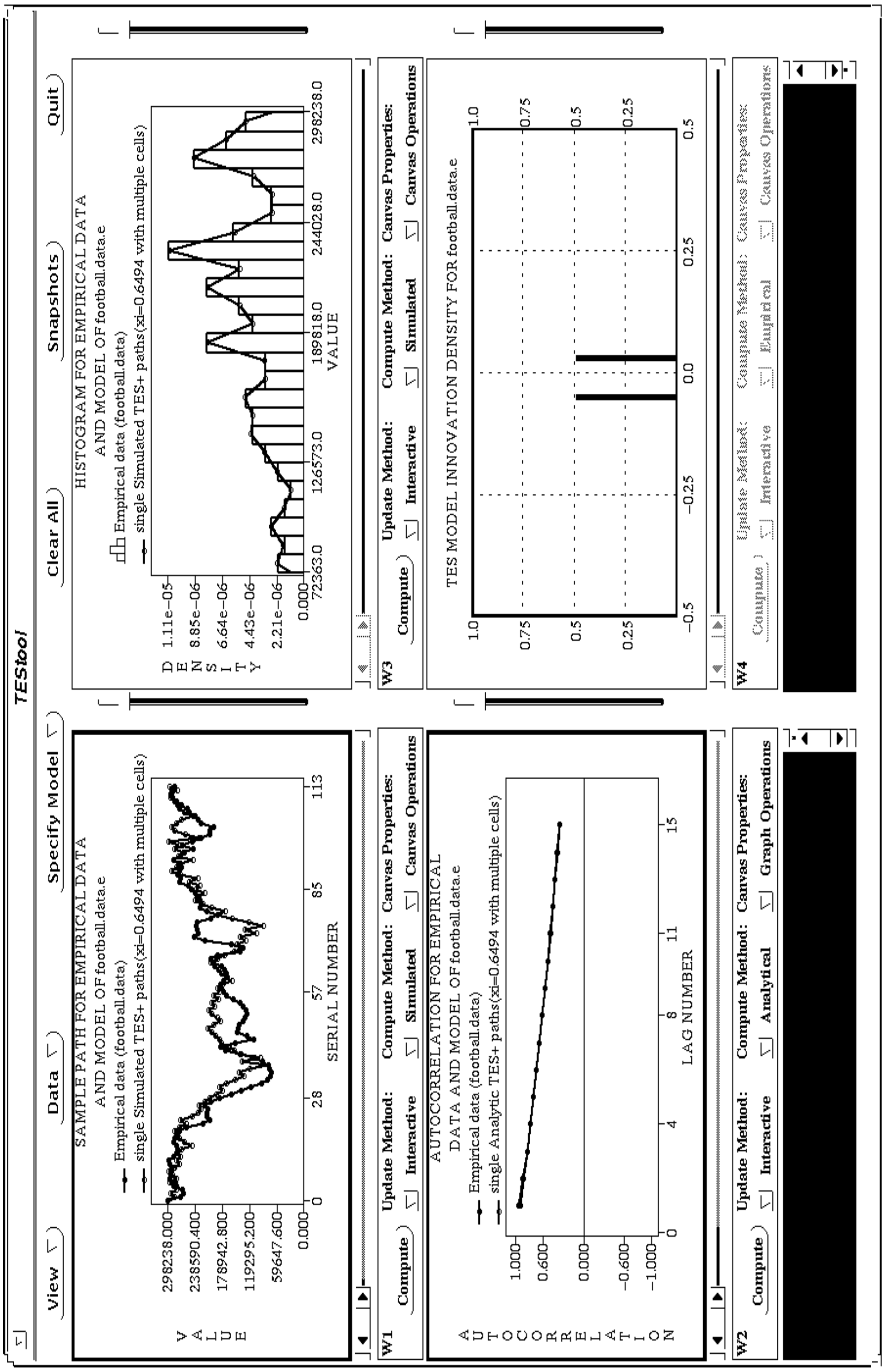

Figure 1: TES Model of DCT-Compressed VBR Video

empirical record of a (random) sequence of encoded (compressed) video frames [58]. The specific data modeled in Figure 1 was a random sequence of frame bit rates, generated by a VBR video sequence of a football scene whose frames were compressed by a variant of the DCT (discrete cosine transform) [82]; see also [77] for a review of compressed video modeling, using the TES methodology. The screen is sub-divided into four canvas areas, designed to display various types of graphics; the graphical user interface is controlled by buttons and menu selections. Each statistical canvas superimposes an empirical statistic with its TES-model counterpart; a legend at the top of these graphs identifies the constituent curves. The upper-left canvas contains the empirical and model sample paths, the latter being generated by a Monte Carlo simulation; the upper-right canvas contains the corresponding histograms; the lower-left canvas contains the empirical autocorrelation function and its numerically-computed model counterpart; and the lower-right canvas contains a joint specification of a TES sign, stitching parameter and an innovation density.

Notice that the empirical time series in Figure 1 (lower-left canvas) exhibits very high short-range autocorrelations, attesting to the bursty nature of this type of encoded video. A simple renewal model would not be appropriate, since it cannot capture burstiness due to temporal dependence. In contrast, the TES model displayed in Figure 1 exhibits general good statistical agreement with the empirical data, in accord with the three fidelity requirements stipulated above; in particular, the histogram fit is very close in the upper-right canvas, as is the autocorrelation fit in the lower-left canvas (the two autocorrelation curves are visually indistinguishable). Such source models can be used to generate synthetic streams of realistic traffic to drive simulations of communications networks.

2.6 SELF-SIMILAR TRAFFIC MODELS

Recent traffic studies from working packet networks [7, 24, 33, 59, 97, 75] have revealed new features of packet traffic that have gone unnoticed in the traffic modeling literature, yet seem to have serious implications for designing, engineering and controlling future high-speed networks. The measured data demonstrate convincingly that packet traffic can be statistically *self-similar* or, more colloquially, *fractal* in nature. Self-similarity means, roughly, that the traffic looks statistically the same over a wide range of time scales, and is related to *long-range dependence* or $1/f$*-noise*. The term "self-similar" was originally coined by Mandelbrot [74] who illustrates that self-similarity is observed in a wide range of physical and mathematical systems and that it is inherent in the study of many irregular or bursty phenomena. This section introduces second-order self-similar processes, discusses some of their characteristic features, and exemplifies several stochastic models that can capture self-similarity in a parsimonious manner.

2.6.1 Motivation and Pictorial Evidence

In order to motivate the use of self-similar processes for traffic modeling purposes, we recall the analysis in [59] of high-resolution Ethernet LAN traffic measurements, and the visually-convincing evidence of its self-similar nature (see Figure 2, *ibid.*) Based on 27 consecutive hours of monitored Ethernet traffic, Figure 2 (a)-(e) depicts a plot sequence of time series of packet counts (i.e., number of packets per time unit) for 5 different choices of time units. Starting with a time unit of 100 seconds in Plot (a), each subsequent plot is obtained from the previous one by increasing the time resolution (decreasing the time unit) by a factor of 10 and then focusing on a randomly chosen subinterval (indicated by a darker shade in each plot). The time unit corresponding to the finest time scale is 10 milliseconds in Plot (e); this plot is "jittered" in order to avoid the visually irritating quantization effect associated with such high resolution, that is, a small amount of noise has been added to the actual observed arrival rates. Observe that all plots are visually rather "similar" to each other, so that arrival rates measured over larger time scale (hours, minutes) are quite indistinguishable to the human eye from those measured over smaller time scales (seconds, milliseconds). In particular, no natural length of a "burst" is discernible: at every time scale ranging from milliseconds to minutes and hours, bursts have the same qualitative appearance. This scale-invariant or "self-similar" feature of Ethernet traffic is drastically different from both conventional telephone traffic and from traditional stochastic models of packet traffic. Such models typically give rise to plots of packet counts which are indistinguishable from white noise after aggregating the original time series over a few hundred milliseconds, as illustrated by the plot sequence (a')-(e') in Figure 2; this sequence was obtained by successive aggregations as in the empirical plot sequence (a)-(e), except that it arose from synthetic traffic generated from a comparable compound Poisson process with the same average packet size and arrival rate as the empirical data. More complicated Markovian arrival processes were observed to give rise to plot sequences, indistinguishable from (a')-(e'). Thus, Figure 2 provides a surprisingly simple method for sharply distinguishing between empirical traffic and standard model-generated traffic, thereby motivating the use of self-similar stochastic processes for traffic modeling purposes.

2.6.2 Second-Order Self-Similar Processes

Let $X = \{X_t\}_{t=0}^{\infty}$ be a *covariance stationary* (*wide-sense stationary*) stochastic process with mean μ_X, variance σ_X^2, and autocorrelation function $\rho_X(k)$. In particular, X is assumed to have an autocorrelation function of the form

$$\rho_X(k) \;\approx\; k^{-\beta}\, L_1(k), \;\text{ as } k \to \infty, \tag{9}$$

where $0 < \beta < 1$ and L_1 is slowly varying at infinity, that is, $\lim_{t\to\infty} L_1(tx)/L_1(t) = 1$, for all $x > 0$; examples of such slowly varying functions are $L_1(t) = \text{const}$ and $L_1(t) = \log(t)$. A stochastic process satisfying relation (9) is said to exhibit *long-range dependence* [6, 16, 93]. In Mandelbrot's terminology [74], long-range dependence is also referred to as the *Joseph Effect*, in reference to the Old Testament figure who had interpreted Pharaoh's dream of the "seven lean cows and the seven fat cows" to mean the "seven fat years and seven lean years" that ancient Egypt was to experience. Intuitively, this notion captures the persistence (or autocorrelation) phenomena observed in many naturally-occurring empirical time series; these manifest themselves in clusters (runs) of consecutive large (or consecutive small) values. More formally, processes with long-range dependence are characterized by an autocorrelation function that decays hyperbolically in the lag. Moreover, it is easy to see that (9) implies $\sum_{k=1}^{\infty} \rho_X(k) = \infty$. This non-summability of the autocorrelations captures the intuition behind long-range dependence, namely, that even though the high-lag autocorrelations are individually small, their cumulative effect is of importance, giving rise to behavior which is markedly different from that of processes with *short-range dependence*; the latter are characterized by geometric decay of the autocorrelation function, that is, $\rho_X(k) \approx r^k$, as $k \to \infty$, for some $0 < r < 1$, resulting in a summable autocorrelation function ($0 < \sum_{k=1}^{\infty} \rho_X(k) < \infty$). In the frequency domain, long-range dependence manifests

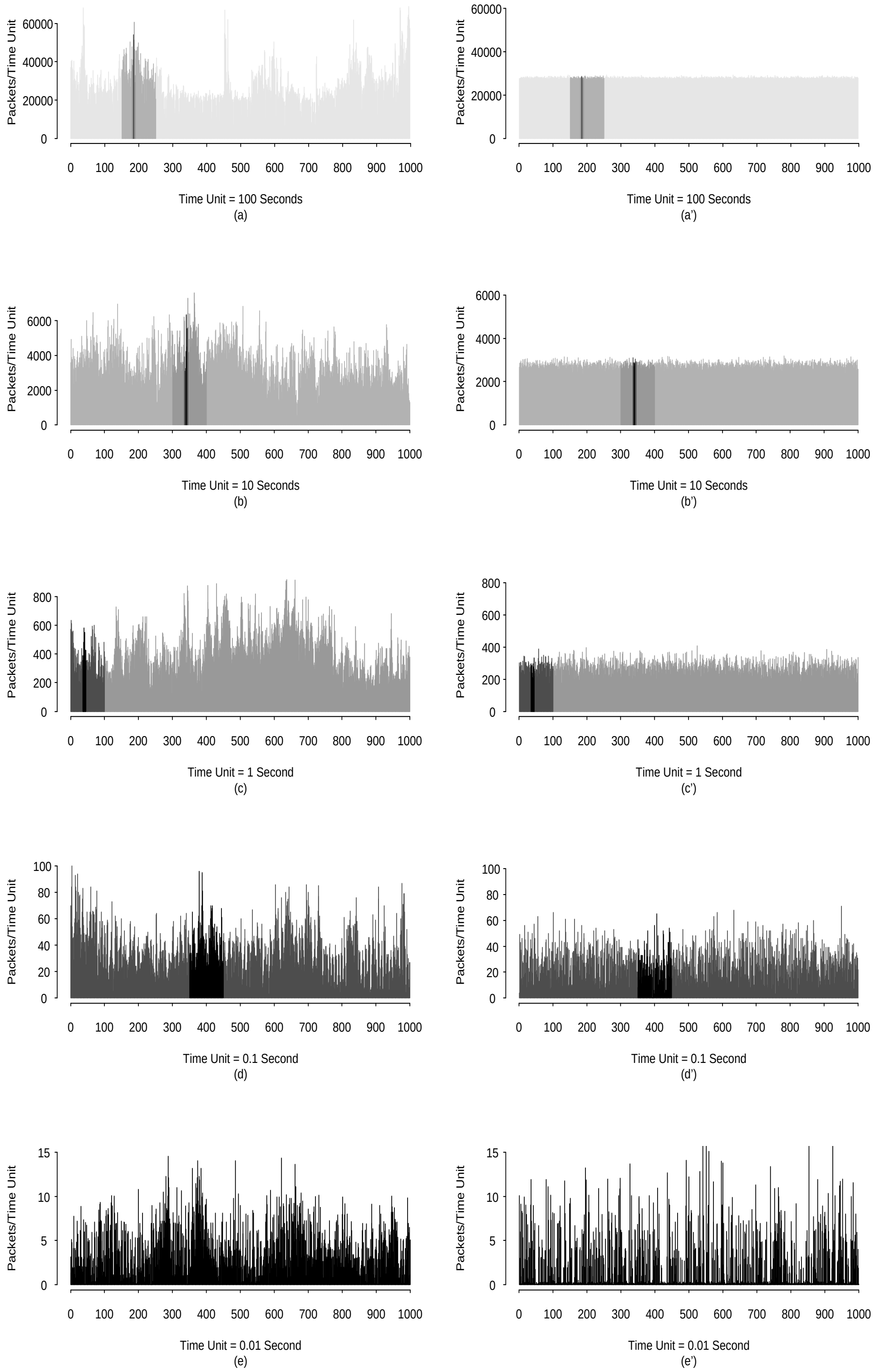

Figure 2: Self-Similar Ethernet Traffic vs. Synthetic Traffic

itself in a spectral density that obeys a power-law near the origin. In fact, under mild regularity on the slowly varying function L_1 in (9), there is long-range dependence in X, if the associated spectral density function, $s_X(\lambda) = \sum_k \rho_X(k)\, e^{ik\lambda}$, satisfies

$$s_X(\lambda) \;\approx\; \lambda^{-\gamma}\, L_2(\lambda), \qquad \text{as } \lambda \to 0, \tag{10}$$

where $0 < \gamma < 1$ and L_2 is slowly varying at 0. Since $s_X(0) = \sum_{k=0}^{\infty} \rho_X(k)$, it follows that long-range dependence is characterized by $s_X(0) = \infty$ (the so-called $1/f$-noise). In contrast, short-range dependence is characterized by $0 \le s_X(0) < \infty$.

For each integer $m \ge 1$, let $X^{(m)} = \{X_k^{(m)}\}_{k=1}^{\infty}$ denote the new covariance stationary time series (with variance $(\sigma_X^{(m)})^2$ and autocorrelation function $\rho_X^{(m)}$)

$$X_k^{(m)} \;=\; \frac{1}{m} \sum_{i=1}^{m} X_{km-m+i}, \tag{11}$$

obtained by averaging the original series, X, over non-overlapping blocks of size m. Self-similarity concepts relate statistical properties of X to their counterparts in $X^{(m)}$, using power laws or invariance, either for all $m \ge 1$, or asymptotically as $m \to \infty$. The process X is called *(exactly) self-similar* with self-similarity parameter $H = 1 - \beta/2$, if the stochastic process $\{m^{1-H} X_k^{(m)}\}_{k=1}^{\infty}$ has the same finite-dimensional distributions as X, for all integer $m \ge 1$. X is called *(exactly) second-order self-similar* with self-similarity parameter $H = 1 - \beta/2$, if the stochastic process $\{m^{1-H} X_k^{(m)}\}_{k=1}^{\infty}$ has the same variance and autocorrelation function as X, for all integer $m \ge 1$. In terms of the aggregated processes $X^{(m)}$, this implies

$$(\sigma_X^{(m)})^2 \;=\; \sigma_X^2\, m^{-\beta}, \tag{12}$$

$$\rho_X^{(m)}(k) \;=\; \rho_X(k) \;=\; \frac{1}{2}\, \delta^2(|k|^{2-\beta}), \tag{13}$$

where $\delta^2(f(k)) = f(k+1) - 2f(k) + f(k-1)$ is the second central difference operator of a sequence $\{f(k)\}$. Finally, X is called *asymptotically (second-order) self-similar* with self-similarity parameter $H = 1 - \beta/2$, if for all $k \ge 0$,

$$\rho_X^{(m)}(k) \;\to\; \frac{1}{2}\, \delta^2(k^{2-\beta}), \qquad \text{as } m \to \infty. \tag{14}$$

Thus, an asymptotically self-similar process has the property that for large m, the corresponding aggregated time series, $X^{(m)}$, have a fixed autocorrelation structure which is solely determined by β; moreover, due to the asymptotic equivalence (for large k) of differencing and differentiating, $\rho_X^{(m)}$ agrees asymptotically with the autocorrelation structure of X, given by (9). Intuitively, the most striking feature of (exactly or asymptotically) self-similar processes is that their aggregated processes possess a nondegenerate autocorrelation structure, as $m \to \infty$. This behavior is in stark contrast to conventional stochastic models, whose aggregated processes tend to second-order pure noise, i.e.,

$$\rho_X^{(m)}(k) \;\to\; 0, \qquad \text{as } m \to \infty \tag{15}$$

for all $k > 0$.

The importance of self-similar processes derives from the fact that they provide elegant models which capture an important empirical law, commonly referred to as *Hurst's law* or the *Hurst effect*, to be reviewed next. Consider a finite sequence of observations $\{X_k\}_{k=1}^n$ with sample mean $\bar{X}(n)$ and sample variance $S^2(n)$, and define $W_k = \sum_{i=1}^k X_i - k\bar{X}(n)$. The *rescaled adjusted range statistic* (*R/S statistic* for short) is given by

$$\frac{R(n)}{S(n)} = \frac{\max(0, W_1, W_2, \ldots, W_n) - \min(0, W_1, W_2, \ldots, W_n)}{S(n)}. \tag{16}$$

Hurst [45] found that many naturally occurring time series appear to be well-modeled by the relation

$$E\left[\frac{R(n)}{S(n)}\right] \approx c\,n^H, \quad \text{as } n \to \infty, \tag{17}$$

where c is a finite positive constant independent on n, and H has typical values around 0.7 [45]. The parameter H is called the *Hurst parameter*. In contrast, if $\{X_k\}$ is a process with short-range dependence, then

$$E\left[\frac{R(n)}{S(n)}\right] \approx d\,n^{0.5}, \quad \text{as } n \to \infty, \tag{18}$$

where d is a finite positive constant, independent of n [70]. The discrepancy between (17) and (18) is generally referred to as the *Hurst effect*.

From a statistical point of view, the most salient feature of self-similar processes is that the variance of the running arithmetic mean decays in the sample size, n, more slowly than the reciprocal of the sample size. More specifically, the variance of the running arithmetic mean of self-similar processes decays like $n^{-\beta}$, for some $0 < \beta < 1$, whereas for processes whose aggregated series converge to second-order pure noise, the same variance decays like n^{-1}. Here, we shall assume, for simplicity, that the slowly varying functions, L_1 in (9) and and L_2 in (10), are asymptotically constant. In fact, it is shown in [16] that a specification of the autocorrelation function satisfying (9) (or equivalently, of the spectral density function satisfying (10)) coincides with a specification of the sequence $\{(\sigma_X^{(m)})^2\}_{m=1}^\infty$, with the property

$$(\sigma_X^{(m)})^2 \approx a\,m^{-\beta}, \quad \text{as } m \to \infty,$$

where a is a finite positive constant independent of m, and $0 < \beta < 1$ is the same as in (9). In fact, β is related to the parameter γ in (10) by $\beta = 1 - \gamma$. In contrast, for covariance stationary processes whose aggregated series $X^{(m)}$ tend to second-order pure noise (i.e., for which (15) holds), the sequence $\{(\sigma_X^{(m)})^2\}_{m=1}^\infty$ satisfies

$$(\sigma_X^{(m)})^2 \approx b\,m^{-1}, \quad \text{as } m \to \infty,$$

where b is a finite positive constant independent of m. The consequences of the slowly-decaying variances, $(\sigma_X^{(m)})^2$, can be disastrous for classical statistical tests and confidence or prediction intervals (see, e.g., [6]), since the usual standard errors (derived for conventional models) are wrong by a factor that tends to infinity in the sample size.

The statistical properties of slowly decaying variances, long-range dependence, and a power-law spectral density are thus seen to be different manifestations of the property that the underlying covariance stationary process, X, is asymptotically or exactly second-order self-similar. Consequently, the problem of testing for self-similarity and estimating it quantitatively can be approached from a number of different angles utilizing both time domain and frequency domain approaches; these include the time-domain R/S analysis and variance analysis of the aggregated processes, and the frequency-domain periodogram analysis. For details on statistical inference for self-similar processes, see [6, 97] and references therein.

We conclude this subsection by pointing out that for continuous-time processes $X = \{X_t\}_{t\geq 0}$ with zero mean and stationary increments, an alternative definition of self-similarity requires that for all $a > 0$,

$$X_{at} \;=\; a^H X_t, \quad t \geq 0, \tag{19}$$

where equality is understood in the sense of equality of the finite-dimensional distributions, and the exponent H is the self-similarity parameter. Nevertheless, we do not use the self-similarity variant (19), because the previous definitions, (13) and (14), have the advantage that they do not obscure the connection with standard time series theory, and they reflect the fact that one is mainly interested in large time scales, m. From a modeling perspective, the crucial point is that both the discrete-time and the continuous-time definitions involve a wide range of time scales. One advantage of (19) in the presence of large data sets is that it permits for a quick heuristic estimation of the self-similarity parameter H from simple plots like those in Figure 2. In fact, a naive inference from the successive plots (a)-(e) in that figure (subtracting the sample mean of X and applying simple statistics, such as range and histogram) yields H-values of about 0.8 for the relation (19). In contrast, Plots (a')-(e') in that figure reveal pure white noise behavior (i.e., $H = 0.5$) for the synthetic traffic model, which result in identical but degenerate autocorrelation structure of the $X^{(m)}$ for $m > 100$.

2.6.3 Stochastic Modeling of Self-Similar Phenomena

Stochastic process with an autocorrelation structure of the form (9) are often viewed as finite approximations of a continuous sum of Gauss-Markov processes; these are often interpreted as suggesting the presence of a multilevel hierarchy of underlying mechanisms which give rise to self-similarity [73]. Some physical examples may be found in references cited by [16]. In general, however, it is very difficult to demonstrate the physical existence of such multilevel hierarchies, or to relate extant mechanisms to self-similar behavior. Consequently, formal mathematical models have been devised to capture self-similarity phenomena, but by and large, these are not amenable to physical interpretation. Two such models, the exactly (second-order) self-similar fractional Gaussian noise process and the asymptotically (second-order) fractional ARIMA process, will be presented next. We also discuss a construction of self-similar models, originally due to Mandelbrot [71] and later extended in [92, 60], which appears to be promising in terms of providing a physical "explanation" for the self-similarity property in high-speed packet traffic (see [97]).

2.6.4 Fractional Gaussian Noise

A *fractional Gaussian noise* [70] is a stationary Gaussian process, $X = \{X_k\}_{k=1}^{\infty}$, with mean μ_X, variance σ_X^2, and autocorrelation function of the form

$$\rho_X(k) \;=\; \frac{1}{2}\,(|k+1|^{2H} - 2\,|k|^{2H} + |k-1|^{2H}), \quad k \geq 1. \tag{20}$$

It can be shown that the asymptotic form of (20) is $\rho_X(k) \approx H(2H-1)|k|^{2H-2}$, as $k \to \infty$ for $0 < H < 1$, and that the attendant aggregated processes, $X^{(m)}$, exhibit long-range dependence in the sense of (9) and satisfy (13). Thus, fractional Gaussian noise is exactly second-order self-similar with self-similarity parameter H, provided $1/2 < H < 1$. In the case $\mu_X = 0$, a fractional Gaussian noise serves as the increment process of *fractional Brownian motion*, i.e., a continuous-time zero-mean Gaussian process, $B_H = \{B_H(t)\}_{t=0}^{\infty}$, with $0 < H < 1$ and autocorrelation function $\rho_X(s,t) = 1/2(|s|^{2H} + |t|^{2H} - |t-s|^{2H})$.

Fractional Gaussian noise and fractional Brownian motion have been particularly popular in hydrological modeling (see, e.g., [72]). Despite its rigid autocorrelation structure, fractional Gaussian noise is often a reasonable first approximation of more complex structures due to the fact that certain long-range dependent processes yield fractional Gaussian noise as limits under a special type of central limit theorem. Methods for estimating the three unknown parameters, μ_X, σ_X^2 and H have been developed.

2.6.5 Fractional ARIMA(p,d,q) Processes

A *fractional ARIMA*(p,d,q) process [37, 44], where p and q are non-negative integers and d is real, is a stochastic sequence, $X = \{X_k\}_{k=1}^{\infty}$, of the form

$$\Phi[B]\,\Delta^d(X_k) \;=\; \Theta[B]\,\epsilon_k, \tag{21}$$

where $\Phi[B] = 1 - \sum_{i=1}^{p} \phi_i\,B^i$, and $\Theta[B] = 1 - \sum_{i=1}^{q} \theta_i\,B^i$ are polynomials in the backward-shift operator $B(X_k) = X_{k-1}$, $\Delta = 1 - B$ denotes the differencing operator, and Δ^d is the fractional differencing operator defined by $\Delta^d = (1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k}(-B)^k$, with $\binom{d}{k}(-1)^k = \dfrac{\Gamma(-d+k)}{\Gamma(-d)\,\Gamma(k+1)}$ and the sequence $\{\epsilon_k\}_{k=0}^{\infty}$ is a white noise process. It is known [37] that for $d \in (-1/2, 1/2)$, X is stationary and invertible, and its autocorrelation function satisfies $\rho_X(k) \approx a k^{2d-1}$ as $k \to \infty$, where a is a finite positive constant independent of k. Moreover, it was shown in [16] that the attendant aggregated time series, $X^{(m)}$, satisfy (14) for $-1/2 < d < 1/2$. Thus, (9) and (14) hold and X is asymptotically second-order self-similar with self-similarity parameter $d + 1/2$, provided $0 < d < 1/2$.

The high-lag autocorrelations of fractional ARIMA(p,d,q) processes are similar to those of the corresponding fractional ARIMA$(0,d,0)$ processes. The latter constitute the simplest and most fundamental of the fractionally differenced ARIMA processes and have an infinite-order autoregressive representation, $\Delta^d(X_k) = \epsilon_k$. The corresponding infinite-order moving average representation, $X_k = \Delta^{-d}(\epsilon_k)$, reveals that ARIMA$(0,d,0)$ processes are obtained by subjecting white noise to fractional differencing of order $-d$. Upon setting $H = d + 1/2$, both the fractional Gaussian noise and the ARIMA$(0,d,0)$ process have autocorrelations which decay asymptotically as k^{2d-1} (with different constants of proportionality).

One of the main advantages of the ARIMA$(0,d,0)$ family over fractional Gaussian noise processes is that the former can be combined with the established class of Box-Jenkins models [9] in a natural way, to yield the family of ARIMA(p,d,q) processes. Fractional ARIMA processes enjoy greater flexible in simultaneous modeling of short-range and long-range behavior than fractional Gaussian noise. The main reason is that a fractional Gaussian noise process is specified by just three parameters, μ, σ^2 and H, which are insufficient to capture a wide range of low-lag autocorrelation structures encountered in practice. ARMA$(1,d,0)$ and ARMA$(0,d,1)$ models possess much higher flexibility [44].

2.6.6 Self-Similarity Through Aggregation

Let $\{I_k\}_{k\geq 0}$ be a sequence of iid integer-valued random variables ("inter-renewal times") with asymptotic tail probabilities obeying the power law

$$P\{I_k \geq t\} \;\approx\; t^{-\alpha}\, h(t), \qquad \text{as } t \to \infty, \tag{22}$$

where $1 < \alpha < 2$ and h is varying slowly at infinity. For example, the stable (Pareto) distribution with parameter $1 < \alpha < 2$ satisfies the "heavy-tail" condition (22). Mandelbrot [74] refers to (22) as the *Noah Effect* or the *infinite variance syndrome*, in reference to the Biblical story of Noah and the Deluge ("Big Flood"). Intuitively, this notion captures the phenomenon that empirical time series can fluctuate far away from their mean value, with non-negligible probability. In addition to $\{I_k\}_{k\geq 0}$, let $\{G_k\}_{k\geq 0}$ be an iid sequence ("rewards"), independent of $\{I_k\}$, with $E[G_k] = 0$ and $E[G_k^2] < \infty$. Consider the stationary (delayed) renewal sequence $\{S_k\}_{k\geq 0}$ defined by $S_k = S_0 + \sum_{j=1}^{k} I_j,\; k \geq 1$, with an appropriately chosen S_0. A discrete-time (renewal) reward process, $W = \{W_k\}_{k\geq 1}$, can then be defined by

$$W_k \;=\; \sum_{n=1}^{k} G_n\, 1_{(S_{n-1},S_n]}(k).$$

Notice that W is stationary in the sense that its finite-dimensional distributions are invariant under time shifts. By aggregating M iid copies, $W^{(1)}, W^{(2)}, \cdots, W^{(M)}$ of W, one obtains the process $W^* = \{W_k^*(M)\}_{k\geq 0}$, given by

$$W_k^*(M) \;=\; \begin{cases} 0, & k = 0 \\ \displaystyle\sum_{n=1}^{k}\sum_{m=1}^{M} W_n^{(m)}, & k > 0 \end{cases}$$

It can be shown [71, 92] that for k and M both large and $k \ll M$, the process W^* behaves like a fractional Brownian motion. More precisely, the process W^*, properly normalized, converges to the integrated version of fractional Gaussian noise, the notion of convergence being that of finite-dimensional distributions. Thus, the increment process of W^* behaves asymptotically like fractional Gaussian noise.

The ability to produce self-similarity by aggregating an increasing number of iid copies of the rather elementary renewal reward process, W, over increasing time periods relies crucially on the heavy tail behavior, (22), of the inter-renewal times I_k. Consequently, the process W can assume the same value, with high probability, over long periods of time. Aggregating a large number of iid copies of W results in an overall sum that approaches a Gaussian distribution; and over long time periods, this procedure introduces significant temporal dependence. These attributes are shared by fractional Brownian motion. A similar construction of asymptotic self-similarity, employing certain AR(1) processes in lieu of the renewal reward process W, may be found in [38].

2.6.7 Parsimonious Modeling

Since empirical data sets are necessarily finite, it is not possible to determine with certainty whether or not the asymptotic relations (9), (13), (14), etc. hold for data records. However, if sufficient data (empirical or simulated) are available and the underlying process is not self-similar (in the sense that the increasingly aggregated series converge to second-order pure noise as per (15)), then the following can be expected to be observed: (i) the autocorrelations will eventually decrease exponentially, (ii) the spectral density function at the origin will eventually prove to be "continuous", (iii) the variances of the aggregated processes will eventually decrease as m^{-1}, and (iv) the rescaled adjusted range will eventually increase as $n^{0.5}$, as per (18).

In practice, statistical inferences of self-similarity from finite sample sizes are generally problematic. For large data sets, it is often possible to investigate "near" asymptotic behavior of statistics, such as the rescaled adjusted range or the variance of the aggregated processes. Moreover, parsimonious modeling then becomes a necessity, due to the large number of parameters needed to fit a conventional model to a data record which is, in fact, self-similar. For example, modeling long-range dependence via ARMA processes is equivalent to approximating a hyperbolically decaying autocorrelation function by a sum of exponentials. Although this is always possible mathematically, the number of requisite parameters will tend to infinity in the sample size; furthermore, physically meaningful interpretations of the ensuing parameters become increasingly difficult. In contrast, long-range dependence can be parsimoniously modeled via a self-similar process by a single parameter — the Hurst parameter, H. Moreover, self-similar processes capture both short-rage and long-range dependence phenomena.

2.7 STOCHASTIC INTENSITIES OF POINT PROCESSES

Stochastic intensity processes provide an alternative characterization of point processes, which utilizes modern martingale theory [20, 46]. The stochastic intensity formulation is primarily of theoretical, not practical, interest. Since most traffic models are formulated as point processes, a brief overview of stochastic intensities is included here for completeness.

Recall the point process, $N = \{N(t)\}_{t=0}^{\infty}$, as defined in Part I. Intuitively, a stochastic intensity process describes the (random) rate at which points (arrivals) occur in N, given the past history of N and possibly additional information. More precisely, the aforementioned past history is represented by a filtration, $\mathcal{F} = \{\mathcal{F}_t\}_{t=0}^{\infty}$, to which N is adapted, i.e., an increasing family of σ-algebras containing the one generated by N, so that the arrival instants, $\{T_n\}_{n=1}^{\infty}$ of N, are $\mathcal{F}_t$-stopping times.

The setting of the general definition of stochastic intensities is a point process, N, adapted to a filtration, $\mathcal{F}$. An $\mathcal{F}$*-intensity* of N is any non-negative, integrable, $\mathcal{F}$-progressive process, $\lambda = \{\lambda(t)\}_{t=0}^{\infty}$, such that

$$E\left[\int_0^{\infty} C(s)\, dN(s)\right] \;=\; E\left[\int_0^{\infty} C(s)\, \lambda(s)\, d(s)\right], \tag{23}$$

for all $\mathcal{F}$-predictable processes, $\{C(t)\}_{t=0}^{\infty}$ (see [10] for details as well as the definitions of progressive measurability and predictability of stochastic processes). The stochastic intensity definition in (23) is not constructive. Nevertheless, it motivates the term "stochastic intensity", via the "change of variable" from $dN(s)$ to $\lambda(s)ds$. A more transparent motivation is gleaned when a stochastic intensity, λ, is right-continuous and bounded. In this case, a constructive definition of λ is given by the almost sure limit,

$$\lambda(t) \;=\; \lim_{h \to 0+} \frac{E[N(t+h) - N(t) | \mathcal{F}_t]}{h}, \quad t \geq 0, \tag{24}$$

where the limit is a time derivative of conditional expected point counts [10]. Note that each conditional expectation on the right-hand side of (24) is usually a random variable (as is the limit), the exception being Poisson processes which give rise to deterministic stochastic intensities.

The martingale approach to point processes [10] affords a rigorous mathematical treatment of stochastic intensities, including a martingale-based characterization of point processes and their stochastic intensities. The martingale approach also provides a unified treatment of dynamical point process systems with a martingale-based calculus. A case in point is the elegant martingale-based result, commonly known as the *change-of-intensities formula* à la Girsanov, roughly, a change-of-measure result for stochastic intensities akin to a Radon-Nikodym change of the underlying probability measure (*ibid.*)

While the martingale approach to point processes and its calculus has been successfully used to tackle a number of interesting problems in queueing theory, their usefulness in stochastic modeling, statistical inference and generation of synthetic arrival streams remains largely an open question. On the other hand, intensity-based modeling of point processes has proved successful in the area of survival analysis, largely due to important contributions made by modern martingale theory to statistical inference in the domain of intensity-based point processes.

3 TELETRAFFIC METHODS

The traffic models surveyed in Section 2 were treated as mathematical objects in their own right. In practice, traffic per se is of interest primarily as an ingredient of server systems, and attention is focused not on traffic statistics, but on queueing statistics such as blocking probabilities and waiting times.

Traditional telephony addressed traffic with a view to engineering, dimensioning and provisioning the telephone network, mainly for voice services. Queueing-based methods were developed over time and put into use to ensure that a prescribed quality of service (QOS) would be satisfied. For instance, POTS (Plain Old Telephone Service) might prescribe a bound on blocking probabilities (known as "grade of service" in traditional telephony), e.g., that at most 1% of dialed calls experience blocking. Even though emerging high-speed communications networks, such as B-ISDN (broadband integrated services digital networks) under ATM (asynchronous transfer mode), are slated to carry new and far more varied traffic than traditional voice and data networks (notably compressed video and images), the methods to be surveyed below will likely be of use in such new communications networks.

The material in Section 3 treats traffic within a queueing framework, with a view to practical teletraffic engineering. While the treatment is mathematical, the focus is on explicit and computable formulae and the computational aspects thereof. Some traffic methods for heavy traffic may be found in Section 3.2. For light-traffic methods, the reader is referred to [8].

3.1 TRAFFIC BURSTINESS

A recurrent theme relating to traffic in broadband networks is the traffic "burstiness" exhibited by key services, such as compressed video, file transfer etc. Intuitively, burstiness is present in a traffic process, if the arrival points, $\{T_n\}$, appear to form visual clusters; equivalently, $\{A_n\}$ tends to give rise to runs of several relatively short interarrival times, followed by a relatively long one. The mathematical underpinning of burstiness is more complex. Two main sources of burstiness are due to the shapes of the marginal distribution and autocorrelation function of $\{A_n\}$. For example, burstiness would be facilitated by a bimodal marginal distribution of $\{A_n\}$, or by short-range autocorrelations in $\{A_n\}$. Strong positive autocorrelations are a particularly major cause of burstiness. Since there seems to be no single widely-accepted notion of burstiness, we shall briefly describe some of the commonly-used mathematical measures which attempt to capture it. Engineers tend to employ very simple measures of burstiness which only use first-order traffic statistics, namely, those associated with interarrival distributions. A typical example is the *peak-to-mean ratio* (PMR) of traffic rate — a very crude measure, which also has the shortcoming of dependence on the interval length utilized for rate measurement. Another example is the coefficient of variation of interarrival times, c_A. Generally, these constitute weak characterizations which only capture very crude aspects of traffic burstiness. More useful measures of burstiness utilize second-order properties of $\{A_n\}$, namely, those associated with its autocorrelation function. Finally, the *Hurst parameter* has been suggested as a measure of burstiness via the concept of self-similarity [59]; see Section 2.6.

3.1.1 The PTC and ISE Traffic Principles

The two burstiness descriptors, indices of dispersion and peakedness, are motivated, respectively, by two traffic principles; the Poisson Traffic Comparison (PTC) principle, and the Infinite-Server Effect (ISE) principle. These principles trace their origin to teletraffic analysis and design, and will be explained next.

The PTC principle compares a traffic-related descriptor (statistic), d_X, of a target traffic process, $\{X_n\}$, to the corresponding descriptor, d_E, of a "comparable" Poisson process, $\{E_n\}$, where the E_n are iid exponential random variables. The descriptor statistic is often related to moments of the underlying process. The PTC principle adopts the view that Poisson processes constitute a class of traffic benchmarks; indeed, these classic traffic processes have a long history in teletraffic theory and practice. Furthermore, "comparability" usually means that the benchmark Poisson process is chosen to have the same rate as the target traffic process. The mathematical comparison is most often performed by taking the target-to-benchmark descriptor ratio. In particular, under the PTC principle, a generic burstiness descriptor, b_X, has the form

$$b_X \;=\; \frac{d_X}{d_E}. \tag{25}$$

Eq. (25) has the additional attraction that PTC-based descriptors of any Poisson process trivially evaluate to unity, as befits a benchmark. The indices of dispersion descriptors of burstiness, treated in Section 3.1.2, are motivated by the PTC principle.

The ISE principle adopts the view that to gauge an aspect of a target traffic stream, $\{X_n\}$, one offers that stream to an infinite-server system and observes the stream effect via a suitable statistic, d_X; infinite-server systems are chosen due to their relative tractability. For example, to characterize burstiness, a suitable description would be the equilibrium moments of the number of customers in the system. The PTC principle is often employed implicitly, by using a normalized form of the description in such a way that it evaluates to unity for any Poisson process. The ISE principle is also used to construct an approximation to a target traffic stream via a mathematically simpler stream. This is attained by equating equilibrium moments of the number of busy servers induced by the two streams. An example is the "three moment match" for creating an interrupted Poisson stream approximation to overflow traffic [57]. The burstiness notion of peakedness, treated in Section 3.1.3, is subject to the ISE principle.

3.1.2 Indices of Dispersion

Consider a stationary, non-negative random sequence, $\{X_n\}_{n=0}^{\infty}$, interpreted as the interarrival times of a traffic process. Let the common distribution function of the X_n be denoted by $F_X(x)$; similarly, denote $\lambda_X = 1/E[X_n]$ (traffic rate), $\sigma_X^2 = Var[X_n]$, and $c_X = \lambda_X\,\sigma_X$. We assume that $0 < \sigma_X < \infty$, and that $\{X_n\}$ is simple, namely, $P\{X_n = 0\} = 0$. The coefficient of variation, c_X, can be used to characterize traffic burstiness, albeit rather weakly, since it utilizes just first-order statistics of $\{X_n\}$ (those pertaining to $F_X(x)$ only). In contrast, the second-order properties are characterized by the autocorrelation function of $\{X_n\}$, given by

$$\rho_X(j) \;=\; \frac{E[X_k\,X_{k+j}] - \lambda_X^{-2}}{\sigma_X^2}, \quad j = 0, 1, \ldots, \tag{26}$$

which measures temporal linear dependence among lagged interarrival times, and will prove to be a key ingredient in the burstiness descriptors to be discussed in the sequel.

The index of dispersion for intervals (IDI) [40], associated with $\{X_n\}$ is a burstiness descriptor defined by

$$J_X(n) \;=\; \frac{Var[\sum_{j=1}^{n} X_j]}{n\,\lambda_X^{-2}}. \tag{27}$$

In order to relate the IDI to the PTC principle, denote $v_X(n) = Var[\sum_{j=1}^{n} X_j]$, and note that for a Poisson process with interarrival times, $\{E_n\}$, of mean $1/\lambda_E$, one has $v_E(n) = n\,\lambda_E^{-2}$. Thus, Eq. (27) is the ratio, $J_X(n) = v_X(n)/v_E(n)$, in adherence to the PTC principle. A simple computation provides the explicit relation of $J_X(n)$ to c_X and ρ_X, namely,

$$J_X(n) \;=\; c_X^2 \left[1 \;+\; 2 \sum_{j=1}^{n-1} \left(1 - \frac{j}{n} \right) \rho_X(j) \right] . \tag{28}$$

Let $\{N(t)\}$ be the equilibrium counting process of arrivals in the interval $(0, t]$, so that $E[N(t)] = \lambda_X t$. A related burstiness descriptor is the index of dispersion for counts (IDC) [40], associated with $\{X_n\}$, and given by

$$I_X(t) \;=\; \frac{Var[N(t)]}{E[N(t)]} \;=\; \frac{Var[N(t)]}{\lambda_X t}, \quad t \ge 0. \tag{29}$$

That the IDC adheres to the PTC principle follows from the fact that the mean and variance of a Poisson counting process are equal.

The limiting indices of dispersion, $J_X = \lim_{n\to\infty} J_X(n)$ and $I_X = \lim_{t\to\infty} I_X(t)$, are of particular interest. It can be shown that [40]

$$J_X \;=\; I_X \;=\; c_X^2 \left[1 \;+\; 2 \sum_{j=1}^{\infty} \rho_X(j) \right] . \tag{30}$$

The indices of dispersion, (27)–(30), are more satisfactory descriptors of traffic variability (burstiness) than the coefficient of variation, c_X, since they take into account the autocorrelation function, $\rho_X(j)$, as well. In fact, for renewal traffic, $J_X(n) \equiv J_X = I_X = c_X^2$. Moreover, these indices of dispersion may be usefully applied to the study of the effect of an offered load on a server system [88, 29, 40].

In practical queueing studies, one uses the indices of dispersion, $J_X(n)$ and $I_X(t)$, and especially the simpler common limit $J_X = I_X$, by estimating their values from empirical observations and then fitting a suitable traffic model, such as a Markov modulated Poisson process (MMPP) which is consistent with the observed index of dispersion [40]. Having thus constructed a model, a queueing system is analyzed with this traffic as the offered load. While this approach has led to some useful results, the index-of-dispersion characterization of traffic burstiness is still fairly weak from a queueing viewpoint. The reason is that indices of dispersion do not consider server systems in their definition. This situation will be remedied in the next section which will introduce a more powerful traffic burstiness characterization.

3.1.3 The Peakedness Functional

Peakedness is a traffic burstiness descriptor, based on the ISE principle. It gauges the "smoothness" of a traffic stream, X, via its effect (as offered traffic) on an iid infinite-server group, i.e., all service times are mutually independent with common service distribution $B(x)$ (usually the exponential distribution) of mean $1/\mu_B$ (μ_B being the service rate). The *peakedness* descriptor of X is the equilibrium variance-to-mean ratio of $S(t)$ — the number of busy servers at time t. R. Wilkinson [96] had introduced the peakedness concept in teletraffic theory in order to characterize the action of superposition of overflow streams when offered to a secondary trunk group. The object was to compute the grade of service (blocking probability) in the context of his "equivalent random method" (see [48] and Section 3.5). A simple and more flexible computational method utilizing peakedness for grade-of-service computation was introduced later by W.S. Hayward (see *ibid.* and Section 3.6.) An advantage of peakedness is that it may be used to approximate various statistics in a single-server system, with or without finite buffers, such as waiting times and blocking probabilities [26, 48].

Let $a = \lambda_X/\mu_B$, and note that Little's Law implies $\lim_{t\uparrow\infty} E[S(t)] = a$. The *peakedness functional*, $z_X[B]$, maps the service time distribution, $B(x)$, to the following scalar [96]

$$z_X[B] \;=\; \lim_{t\uparrow\infty} \frac{Var[S(t)]}{E[S(t)]} \;=\; \frac{\sigma_S^2}{a}, \tag{31}$$

where σ_S^2 is the time-equilibrium variance of the number of busy servers.

For any service distribution, $B(x)$, define $B_0(x) = B(x/\mu_B)$, so that $B_0(x)$ has unit service rate. Define further $B_\mu(x) = B_0(\mu x)$ to be a family of service time distribution functions, parameterized by $\mu > 0$. The peakedness function associated with this parametric family is

$$z_{X,B_0}(\mu) \;=\; z_X[B_\mu]. \tag{32}$$

Let $M_X(t)$ be the expectation function of the traffic process, $\{X_n\}$, that is, the mean number of arrivals in $(0,t]$ when the origin is an arrival point; thus, $M_X(t)$ is the analog of the renewal function in renewal theory. Let $B^c(x)$ designate the complementary service time distribution, that is, $B^c(x) = 1 - B(x)$, and let $\psi_B(x)$ denote the correlation function

$$\psi_B(t) \;=\; \int_0^\infty B^c(u)\, B^c(t+u)\, du, \tag{33}$$

of $B^c(x)$. Then one has [26]

$$z_{X,B_0}(\mu) \;=\; 1 \;-\; \frac{\lambda}{\mu} \;+\; 2\mu \int_0^\infty \psi_B(t)\, dM_X(t). \tag{34}$$

Furthermore, if the density of the expectation function, $m_X(t) = dM_X(t)/dt$, exists then one may also write

$$z_{X,B_0}(\mu) \;=\; 1 \;-\; \frac{\lambda}{\mu} \;+\; 2\mu \int_0^\infty \psi_B(t)\, m_X(t)\, dt. \tag{35}$$

An important special case occurs for exponential service time distributions, $B(x) = 1 - e^{-\mu x}$. The notation $z_{X,exp}(\mu)$ and $\psi_{exp}(x)$ will be used for this case. Since

$$\psi_{exp}(y) \;=\; \frac{1}{2\,\mu}\, e^{-\mu y}, \tag{36}$$

it follows that

$$z_{X,exp}(\mu) \;=\; 1 \;-\; \frac{\lambda_X}{\mu} \;+\; \int_0^\infty e^{-\mu t}\, m_X(t)\, dt. \tag{37}$$

Letting $\tilde{f}(s) = \int_0^\infty e^{-sx}\, f(x)\, dx$ denote the Laplace transform of a function, $f(x)$, one has

$$z_{X,exp}(\mu) \;=\; 1 \;-\; \frac{\lambda_X}{\mu} \;+\; \tilde{m}_X(\mu)\;. \tag{38}$$

Eq. (38) has an important consequence. It implies that when $\tilde{m}(s)$ does not depend on μ, then knowledge of $z_{X,exp}(\mu)$ suffices to determine $z_{X,B}(\mu)$ for *any* service distribution $B(x)$. To see that, observe that from $z_{X,exp}(\mu)$, one can determine $\tilde{m}_X(\mu)$, from which $m_X(t)$ is obtained by inversion. Eq. (35) can now be used to obtain $z_{X,B}(\mu)$ with the aid of $m_X(t)$.

The transformation theory of peakedness (from one service time distribution to another) may be exhibited in a more satisfactory and illuminating form via the Mellin transform [48]. The Mellin transform, $\bar{f}(s)$ of a function $f(x)$, is defined by

$$\bar{f}(s) \;=\; \int_0^\infty x^{s-1}\, f(x)\, dx,$$

and exists in a strip or half-plane [22] of the complex s-plane. Note that $\psi_B(x) = \psi_{B_0}(\mu x)/\mu$ and $\int_0^\infty \psi_{B_0}(\mu x)\, dx \;=\; 1/(2\mu)$. Introducing the notation, $\zeta_{X,B}(\mu) \;=\; z_{X,B_0}(\mu) - 1$ and $h_X(x) = m_X(x) - \lambda_X$, it follows that Eq. (35) is equivalent to

$$\zeta_{X,B_0}(\mu) \;=\; 2\int_0^\infty \psi_{B_0}(\mu x)\, h_X(x)\, dx. \tag{39}$$

Applying the Mellin transform to (39) results in

$$\bar{\zeta}_{X,B_0}(s) \;=\; 2\bar{\psi}_{B_0}(s)\, \bar{h}_X(1-s), \tag{40}$$

which clearly separates the individual effect of the service distribution (via $\bar{\psi}_{B_0}(s)$) and the traffic stream (via $\bar{h}_X(1-s)$) on $\bar{\zeta}_{X,B_0}(s)$.

Suppose now that it is desired to switch from service distribution F to service distribution G, thereby transforming $z_{X,F_0}(\mu)$ to $z_{X,G_0}(\mu)$. From (40), one has

$$\begin{aligned}\bar{\zeta}_{X,F_0}(s) &= 2\bar{\psi}_{F_0}(s)\, \bar{h}_X(1-s),\\ \bar{\zeta}_{X,G_0}(s) &= 2\bar{\psi}_{G_0}(s)\, \bar{h}_X(1-s),\end{aligned}$$

whence, $\bar{\zeta}_{X,F_0}$ and $\bar{\zeta}_{X,G_0}$ are simply related by

$$\frac{\bar{\zeta}_{X,G_0}(s)}{\bar{\zeta}_{X,F_0}(s)} \;=\; \frac{\bar{\psi}_{G_0}(s)}{\bar{\psi}_{F_0}(s)}. \tag{41}$$

Thus, Eq. (41) permits the calculation of traffic peakedness relative to a new service distribution, based on the corresponding peakedness relative to a given service distribution, provided the Mellin transform exists at least as an extension to a singular integral [65].

Suppose next that $z_{X,F_0}(0)$ is known and possibly some of its derivatives at the origin for some service distribution F. From (39), one deduces that the n-th derivative of ζ_{X,F_0} at the origin is given by

$$\zeta^{(n)}_{X,F_0}(0) \;=\; 2\,\psi^{(n)}_{F_0}(0) \int_0^\infty x^n\, h(x)\, dx, \tag{42}$$

provided the differentiations are permissible within the integral. Combining (42) with the corresponding formula for $\zeta_{X,G_0}(\mu)$, and denoting

$$R_n \;=\; \frac{\psi_{F_0}^{(n)}(0)}{\psi_{G_0}^{(n)}(0)}, \tag{43}$$

one obtains

$$\zeta_{G_0}^{(n)}(0) \;=\; R_n\, \zeta_{F_0}^{(n)}(0). \tag{44}$$

Eq. (44) can be used as a transformation formula for the derivatives of the peakedness at $\mu = 0$, analogously to Eq. (41). For the evaluation of the ratios, R_n in (43), observe that for any service distribution, B, one has

$$\psi_{B_0}^{(n)}(0) \;=\; \int_0^\infty B_0^c(y)\,(B_0^c)^{(n)}(y)\,dy, \tag{45}$$

provided the interchange of differentiation and integration is valid. Applying Eq. (45) separately to the numerator and denominator of R_n, one obtains the particular cases

$$R_0 \;=\; \frac{\int_0^\infty G_0^c(y)^2\,dy}{\int_0^\infty F_0^c(y)^2\,dy}, \quad R_1(y) \;\equiv\; 1. \tag{46}$$

From Eq. (46) follows the important peakedness formula

$$z_{X,G_0}(0) \;=\; 1 \;+\; \frac{\int_0^\infty G_0^c(y)^2\,dy}{\int_0^\infty F_0^c(y)^2\,dy}\,[z_{X,F_0}(0) - 1]. \tag{47}$$

Interestingly, Eq (46) implies the remarkable conclusion that $z'_{X,F_0}(0)$ is invariant under change of service distribution; of course, this does not apply to the higher derivatives.

Practical applications of the peakedness function $z_B(\mu)$ to queueing problems are found in [48, 26, 32].

3.2 THE ERLANG B LOSS FUNCTION

The Erlang B loss function is used to obtain the loss rate of traffic offered to a finite-server group. It is traditional to express traffic units in erlang units, namely, the mean number of arrivals during a time interval whose length is the mean service time.

The Erlang B setting is an $M/G/n/n$ queue. Here, a Poisson arrival stream of rate λ is offered to an iid group of n servers, each with service distribution $B(x)$ of rate μ (i.e., with mean service time $1/\mu$); the offered load is denoted by $a = \lambda/\mu$ and is expressed in erlang units. Let S be the number of busy servers in equilibrium, with probability distribution [56]

$$P\{S = k\} \;=\; \frac{a^k/k!}{\sum_{j=0}^n a^j/j!}, \quad k = 0, 1, \ldots$$

The PASTA property ensures that the distribution of S coincides with its counterpart, embedded at arrival points (see, e.g., [78]). In particular, the probability that all servers are busy is

$$B(n,a) \;=\; \frac{a^n/n!}{\sum_{j=0}^{n} a^j/j!}. \tag{48}$$

Eq. (48) is called the *Erlang B loss function* (Erlang B, for short), and is a function of the server group size and offered load. Thus, the equilibrium mean number of busy servers, called the *carried load*, is $a'(n,a) = a[1-B(n,a)]$, and the *overflow rate of blocked customers* (in erlangs) is $aB(n,a)$.

Suppose that the servers are numbered 1 to n and the entire arrival stream is offered to server 1, the overflow then offered to server 2 and so on; this discipline is called an *ordered hunt*. Thus, the load carried by server j is

$$\ell(j,a) \;=\; a[B(j-1,a) - B(j,a)], \quad 1 \le j \le n. \tag{49}$$

When $j = n$, then Eq. (49) is called the *load on the last server*. Its importance is due to the economic aspects of sizing server groups [56]. A useful identity, relating a, $a'(n,a)$ and $\ell(n,a)$, is

$$a \;=\; a'(n,a)\left[1 + \frac{\ell(n,a)}{n - a'(n,a)}\right]. \tag{50}$$

The variance of S is

$$\sigma_S^2 \;=\; -a\,n \;+\; [a+n+1]\,a'(n,a) \;-\; (a'(n,a))^2 \;=\; [1-\ell(n,a)]\,a'(n,a).$$

Practical applications of the Erlang B theory require knowledge of sensitivities, approximations and asymptotic relations for $B(n,a)$, $\ell(n,a)$ etc. To this end, let $\rho = a/n$ be the *offered load per server*, and let $\eta(n,a) = a'(n,a)/n$ be the *carried load per server* (also known as the *efficiency of the server group*). The following are useful identities,

$$\frac{\partial B(n,a)}{\partial a} \;=\; \left[\frac{n}{a} - 1 + B(n,a)\right] B(n,a) \tag{51}$$

$$\frac{\partial \ell(n,a)}{\partial \rho} \;=\; n\left[1 - \frac{\eta(n,a)}{\rho}\right] \frac{n\,(n\rho - \eta(n,a))(1-\eta(n,a)) - \eta(n,a)^2}{\eta(n,a)^2} \tag{52}$$

$$\frac{\partial \ell(n,a)}{\partial \eta(n,a)} \;=\; \frac{\ell(n,a)}{\eta(n,a)}\,\frac{n\,(1-\eta(n,a))^2 - \eta(n,a) + \ell(n,a)}{(1-\ell(n,a))(1-\eta(n,a))} \tag{53}$$

$$\frac{\partial a'(n,a)}{\partial a} \;=\; \frac{\sigma_S^2}{a} \tag{54}$$

A numerically convenient formula for calculating $B(n,a)$ is via

$$B(n,a)^{-1} \;=\; \sum_{j=0}^{\infty} n^{(j)}\,a^{-j}, \tag{55}$$

where $n^{(j)}$ is the descending factorial, given by

$$n^{(j)} \;=\; \begin{cases} 1, & j = 0 \\ n(n-1)\ldots(n-j+1), & j \ge 1 \end{cases}$$

Eq. (55) is especially useful when $a > n$, since then the series may be conveniently truncated to achieve a given accuracy [47]. Substitution of the identity $a^{-j} = a \int_0^\infty e^{-ay} \frac{y^j}{j!}\, dy,\ a > 0,\ j \geq 0$, into (55) yields Fortet's integral representation

$$B(n,a)^{-1} \;=\; a \int_0^\infty e^{-ay} (1+y)^n \, dy \;=\; \int_0^\infty e^{-y} \left(1 + \frac{y}{a}\right)^n dy. \tag{56}$$

In order to obtain the sensitivity of $B(n,a)$ with respect to n, one extends its representation in (56) to complex n and a, by means of the interpolatory analytic function

$$B(x,a)^{-1} \;=\; \int_0^\infty e^{-y} \left(1 + \frac{y}{a}\right)^x dy, \quad Re[a] > 0. \tag{57}$$

Furthermore, this permits the evaluation of the loss function for non-integral number of servers in practical traffic-related problems. The corresponding extension of Eq. (55) is the asymptotic expansion, as $a \to \infty$ (see [47])

$$B(x,a)^{-1} \;\sim\; \sum_{j=0}^{\infty} x^{(j)} a^{-j}, \quad |\arg a| < \pi. \tag{58}$$

Eqs. (50) and (52) - (54) may be similarly extended. These extensions are computationally convenient, since the error does not exceed the absolute value of the first neglected term. Furthermore, an integration by parts of (57) yields the useful recurrence relation

$$B(x+1,a)^{-1} \;=\; \frac{x+1}{a} B(x,a)^{-1} \;+\; 1. \tag{59}$$

For integral $x = n$, $B(n,a)$ can be conveniently computed via (59) from the initial value $B(0,a) = 1$.

The derivative, $\partial B(x,a)/\partial a$, yields the sensitivity with respect to the offered load, a. It may be obtained from (57) and simplified to

$$\frac{\partial B(x,a)}{\partial a} \;=\; \left[\frac{x}{a} \;-\; 1 \;+\; B(x,a)\right] B(x,a).$$

Using again (57), one has

$$\frac{\partial B(x,a)}{\partial x} \;=\; B^2(x,a) \int_0^\infty e^{-y} \left[1 + \frac{y}{a}\right]^x \ln\left(1 + \frac{y}{a}\right) dy. \tag{60}$$

The evaluation of (57) and (60) can be achieved by means of quadrature theory. The basic goal is the evaluation of the integral $I = \int_0^\infty e^{-y} f(y) dy$. It is known that the trapezoidal rule of quadrature is very accurate when the integral has high-order osculation at the endpoints of integration [17]. Using the transformation $w = \ln y$, one has

$$I \;=\; \int_{-\infty}^{\infty} e^{w - e^w} f(e^w)\, dw. \tag{61}$$

For functions of the form $f(y) = O\Big(\frac{e^y}{y(\ln y)^2}\Big)$, there is sufficient osculation at $-\infty$ and ∞ for the use of the trapezoidal rule. Thus, applying the rule to (61) with span $h > 0$,

$$I \;\simeq\; h \sum_{j=-\infty}^{\infty} e^{jh - e^{jh}} f(e^{jh}). \tag{62}$$

The quadrature rule (62) can be applied to (57) and (60) to obtain $B(x,a)$ and $\partial B(x,a)/\partial x$, respectively. It should be noted that the requisite computational effort is essentially independent of x and a; this may be contrasted with the corresponding effort associated with (58) and (59).

An alternate computational method is based on the extension of the Poisson distribution to continuous x via the function $\psi(x,a) = e^{-a}\dfrac{a^x}{\Gamma(x+1)}$, $x > -1$. From [47], $B(x,a)$ is well approximated by $\psi(x,a)$ when a/x is small, the exact relation being

$$B(x,a)^{-1} \;=\; \psi(x,a)^{-1} \;-\; \sum_{j=1}^{\infty} \frac{a^j}{(x+1)\cdots(x+j)}. \tag{63}$$

The series is rapidly convergent for $x > -1$ and $a/x \leq 1$.

The above computations are accurate but time consuming. When real-time computations are called for, it is desirable to trade accuracy for time and to construct fast approximate formulas for $\partial B(x,a)/\partial x$, $a'(n,a)$, $\ell(x,a)$, etc.

From (57), one deduces that $B(x,a)^{-1}$ and $[aB(x,a)]^{-1}$ are log-convex functions of x and a, respectively, for $a > 0$ and all x [47]; also, $B(x,a)$ is convex in x for $a > 0$, $x > 0$ [52]. The log-convexity of $B(x,a)^{-1}$ or the convexity of $B(x,a)$ imply the inequality $\ell(x,a) \leq \eta(x,a)$. Substituting this upper bound into (50), which remains valid with n replaced by the continuous variable x, one gets

$$a \;\leq\; a'(x,a)\left[1 + \frac{a'(x,a)}{x\,[x - a'(x,a)]}\right],$$

which is a useful first-cut estimate of the offered load, a.

Denote $\alpha = (x+1)/a + B(x,a)$. Then Jensen's inequality [41] applied to (57) implies the lower bound

$$\frac{\partial B(x,a)}{\partial x} \;\geq\; -B(x,a)\,\ln\alpha. \tag{64}$$

A simple, useful approximation for $\partial B(x,a)/\partial x$ may be obtained by the following consideration. The Euler-Maclaurin expansion [89] applied to a functional equation of the form $f(x+1) - f(x) = g(x)$ implies

$$f'(x) \;\simeq\; g(x) - \frac{1}{2}\,g'(x). \tag{65}$$

To apply this to obtaining $\partial B/\partial x$, take

$$f(x) \;=\; \ln B(x,a), \tag{66}$$

whence (59) becomes

$$f(x+1) - f(x) \;=\; -\ln\alpha.$$

Relation (65) now provides the approximation (see, e.g., [48])

$$B(x,a)^{-1}\,\frac{\partial B(x,a)}{\partial x} \;\simeq\; -\frac{\ln\alpha - 1/(2a\alpha)}{1 - B(x,a)/(2\alpha)}.$$

Interpolation plays a useful role in some applications. For example, $B(x,a)$ may be measured in practice at integral values of x, yet in some calculations (e.g., the equivalent random method), the value of $B(x, a)$ is needed at a non-integral value of x. A useful approximation is obtained by applying quadratic interpolation to $f(x)$ in (66). Recalling that $\lfloor x \rfloor$ denotes the integral part of x and $\langle x \rangle = x - \lfloor x \rfloor$ its fractional part, one has

$$B(x,a) \simeq B(n,a)^{1-\langle x \rangle} B(n+1,a)^{\langle x \rangle} \left[\frac{B(n+1,a)^2}{B(n,a)\, B(n+2,a)} \right]^{\langle x \rangle (1-\langle x \rangle)/2}, \tag{67}$$

the worst error of (67) occurring at $\langle x \rangle = 0.5$.

An asymptotic formula for $B(x, a)$ (as $x \to \infty$), when a is in the vicinity of x (the *medium traffic condition*), is given by

$$B\left(x, x + c\sqrt{x}\right) \sim \sum_{j=0}^{\infty} a_j(c)\, x^{-\frac{j-1}{2}}, \quad |\arg x| < \frac{\pi}{2},\ c \text{ real}, \tag{68}$$

where the coefficients $a_j(c)$ are given by

$$a_j(c) = \int_0^{\infty} e^{-(\frac{1}{2}v^2 + cv)}\, b_j(v, c)\, dv,$$

with the $b_j(v, c)$ defined by the relation

$$e^{\frac{1}{2}v^2 - v\sqrt{x}} \left[1 + \frac{v}{\sqrt{x}}\right]^x (\sqrt{x} + c) = \sum_{j=0}^{\infty} b_j(v, c)\, x^{-\frac{j-1}{2}}.$$

The first few coefficients are

$$\begin{aligned}
a_0(c) &= e^{\frac{1}{2}c^2} \int_c^{\infty} e^{-\frac{1}{2}u^2} du, \\
a_1(c) &= \frac{2}{3} + \frac{1}{3}c^2 - \frac{1}{3}c^3\, a_0(c), \\
a_2(c) &= -\frac{1}{18}c^5 - \frac{7}{36}c^3 + \frac{1}{12}c + \left[\frac{1}{18}c^5 + \frac{1}{4}c^4 + \frac{1}{12}\right) a_0(c).
\end{aligned}$$

The special case $c = 0$ yields the useful asymptotic relation as $x \to \infty$,

$$B(x,x)^{-1} \sim \sqrt{\frac{\pi x}{2}} + \frac{2}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2x}}, \quad |\arg x| < \frac{\pi}{2}. \tag{69}$$

Eq. (69) may be used in convenient approximate computations of $a'(n, a)$, $\ell(n, a)$ for the case of medium traffic. In particular, consider the load $\ell^*(a') = \ell(x, a)$, carried by the virtual server x, as a function of the carried load, $a' = a'(x, a)$. Let $a'_0 = a'(x, x)$ denote the carried load at the operating point, (x, x). Then the value of $\ell^*(a')$, resulting from changing the carried load away from the operating point, a'_0, can then be approximated by the two-term Taylor expansion of ℓ^* about a'_0, namely,

$$\ell^*(a') \simeq \ell^*(a'_0) + [a' - a'_0] \frac{d\ell^*(a'_0)}{da'}.$$

where, in view of the relation, $\dfrac{d\ell^*(a')}{da'} = \dfrac{1}{x}\dfrac{d\ell(n,a)}{d\eta}$, one obtains the derivative $\dfrac{d\ell^*(a')}{da'}$ from (53) with the aid of (69). A more convenient form of (49) for $\ell(x, a)$ is

$$\ell(x,a) = \left[\frac{x}{1 - B(x,a)} - a\right] B(x,a),$$

from which $\ell(x,x)$ is readily obtained, while a more accurate asymptotic version of (69), as $x \to \infty$, is

$$B(x,x)^{-1} \sim \sqrt{\frac{\pi x}{2}} + \frac{2}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2x}} - \frac{4}{135x}, \quad |\arg x| < \frac{\pi}{2}.$$

The approximation of $B(x,cx)$, for $c > 1$, applies to the case of *heavy traffic* in large server systems. From [47], one has the asymptotic expansion, as $x \to \infty$,

$$B(x,cx)^{-1} \sim \sum_{k=0}^{\infty} g_k(c)\, x^{-k}, \quad |\arg x| < \frac{\pi}{2},\ c > 1,$$

where the coefficients $g_k(c)$, $k = 0, 1, \ldots$, are given by

$$g_k(c) = \left[\frac{c}{c-1}\frac{d}{dc}\right]^k \frac{c}{c-1},$$

implying, in particular,

$$B(x,cx)^{-1} \sim \frac{c}{c-1} - \frac{c}{(c-1)^3}x^{-1} + \frac{2c^2+c}{(c-1)^5}x^{-2}. \tag{70}$$

Eq. (70) may be contrasted with the exact formula (63) and the light-traffic Poisson approximation derived from it, as well as to the asymptotic formula (58) which approximates heavy traffic with a fixed number of servers.

3.3 THE ERLANG C DELAY FUNCTION

The Erlang C setting is an $M/M/n/\infty$ queue. Here, a Poisson arrival stream of rate λ is offered to an infinite FIFO buffer served by n iid servers with exponential service time distribution $B(x) = 1 - e^{-\mu x}$. The offered load is $a = \lambda/\mu$, and the offered load per server is $\rho = a/n < 1$. Since no customers are lost, the carried load is $a'(n,a) = a$.

Let W be the waiting time in the queue (if any) of an arriving customer in equilibrium. The probability, $C(n,a) = P\{W > 0\}$, that an arriving customer finds all servers busy (and, therefore, must wait in line) is given by

$$C(n,a) = \frac{B(n,a)}{1 - \rho\,[1 - B(n,a)]}. \tag{71}$$

Eq. (71) is called the *Erlang C delay function* (Erlang C, for short), and is a function of the server group size and offered load. For the distribution of W, one has

$$\begin{aligned} P\{W > t\} &= C(n,a)\, e^{-(1-\rho)n\mu t}, \\ P\{W > t | W > 0\} &= e^{-(1-\rho)n\mu t}, \end{aligned}$$

and for its mean and variance,

$$\begin{aligned} E[W] &= \frac{C(n,a)}{[1-\rho]\,n\,\mu}, \\ Var[W] &= \frac{1 - [1 - C(n,a)]^2}{([1-\rho]\,n\,\mu)^2}. \end{aligned}$$

Let S be the number of customers in the system in equilibrium, and let $P_j = P\{S = j\}$ be the corresponding equilibrium time probabilities. Then,

$$P_j = \begin{cases} n!\, a^{-n} \, [1-\rho] \, C(n,a), & j = 0 \\ P_0 \, \dfrac{a^j}{j!}, & 1 \le j \le n-1 \\ P_0 \, \dfrac{a^j}{n! \, n^{j-n}}, & j \ge n \end{cases}$$

and

$$E[S] = P_0 \, \frac{a^n}{n!} \left[a \, [B(n,a)^{-1} - 1] \; - \; n \; + \; \frac{1}{[1-\rho]^2} \right].$$

Finally, let Q be the number of customers in the buffer (excluding customers in service) in equilibrium. Then,

$$\begin{aligned} P\{Q = j, W > 0\} &= [1-\rho] \, \rho^j \, C(n,a), \\ P\{Q = j | W > 0\} &= [1-\rho] \, \rho^j. \end{aligned}$$

All of the above formulae may be immediately generalized to a continuous number of servers. One need only replace $n!$ by $\Gamma(n+1)$ and use the generalization of $B(n,a)$.

3.4 THE ENGSET DELAY MODEL

The Engset model is a closed system consisting of a finite number, n, of independent sources (customers) and a finite number, s, of iid exponential servers [14]. Each source goes through cycles of service as follows. Initially a source is idle (neither in service nor waiting for service). After an exponentially-distributed time with parameter γ, the source places a service request and is considered busy. All busy sources waiting for a server to become available wait in a FIFO queue and eventually seize a server for an exponentially-distributed service time of rate μ. On service completion, the source becomes idle, and the cycle starts over again.

Let $P_j(n,s)$ be the equilibrium probability that j sources are busy. Then,

$$P_j(n,s) = \begin{cases} \dfrac{1}{\sum_{j=0}^{n} \pi_j}, & j = 0 \\ \\ \pi_j \, P_0(n,s), & j > 0 \end{cases} \tag{72}$$

where, letting $\hat{a} = \gamma/\mu$,

$$\pi_j = \begin{cases} \binom{n}{j} \, \hat{a}^j, & 0 \le j \le s \\ \\ \dfrac{n^{(j)}}{s! \, s^{j-s}} \, \hat{a}^j, & s+1 \le j \le n \end{cases} \tag{73}$$

Let $\bar{L}$ be the mean number of busy sources, and $\bar{I}$ the mean number of idle sources. Then, $\bar{L} = \sum_{j=1}^{n} j \, P_j(n,s)$, and $\bar{I} = n - \bar{L}$. Since $\hat{a}$ is the offered load per idle source, it follows that the total offered load is

$$a = \hat{a} \, \bar{I}$$

and the total request rate is

$$\lambda \;=\; a\,\mu.$$

Let $\bar{J}$ be the equilibrium mean sojourn time, and $\bar{W}$ the equilibrium mean waiting time. Then these means are related by $\bar{W} = \bar{J} - 1/\mu$, and from Little's law $\bar{J} = \bar{L}/\lambda$. Since λ/n is the request rate per source, it follows that $n\lambda^{-1}$ is the mean time between requests. Hence,

$$\bar{W} \;+\; \frac{1}{\mu} \;+\; \frac{1}{\gamma} \;=\; \frac{n}{\lambda}.$$

Let $\Pi_j(n,s)$ be the equilibrium probability that j sources are busy at a request arrival instant. The relation between the (time average) probabilities, P_j, and the corresponding (customer-average) probabilities, Π_j, is given by

$$\Pi_j(n,s) \;=\; P_j(n-1,s).$$

This formula is an instance of ASTA (Arrivals See Time Averages), provided the arriving customer is excluded (see, e.g., [78]). The distribution of the waiting time, W, is given by

$$P\{W > t\} \;=\; \Pi_0(n,s)\,\frac{(n-1)!\,\hat{a}^s}{s!}\left(\frac{\hat{a}}{s}\right)^{n-s-1} e^{s/\hat{a}-\phi(t)} \sum_{j=0}^{n-s-1} \frac{[\phi(t)]^j}{j!},$$

where $\phi(t) = s\,\mu\,[t + 1/\gamma]$.

For a single-server Engset model ($s = 1$), Eqs. (73) and (72) can be simplified, using

$$P_0(n,1) \;=\; B(n,1/\hat{a}), \quad a \;=\; 1 - B(n,1/\hat{a}). \tag{74}$$

Furthermore, letting $c = n\gamma$,

$$\bar{J} \;=\; n\left[\frac{1}{\lambda} \;-\; \frac{1}{c}\right] \;=\; n\left[\frac{1/\mu}{1 - B(n,\mu n/c)} \;-\; \frac{1}{c}\right]. \tag{75}$$

Convenient asymptotic formulas (as $n \to \infty$) for Eqs. (74) and (75) are derived from the asymptotic results of Section 3.2 for the special case $s = 1$. These are summarized below for various ranges of μ.

For $\mu > c$, $\quad P_0(n,1) \;\sim\; 1 \;-\; \dfrac{c}{\mu} \;+\; \dfrac{c^2}{n\,\mu\,(\mu - c)}, \quad \bar{J} \;\sim\; \dfrac{1}{\mu - c}$ [from (70)].

For $\mu = c$, $\quad P_0(n,1) \;\sim\; \sqrt{\dfrac{2}{\pi n}} \;-\; \dfrac{4}{3\pi n}, \quad \bar{J} \;\sim\; \dfrac{1}{c}\left(\sqrt{\dfrac{2n}{\pi}} \;+\; \dfrac{2}{3\pi}\right)$ [from (69)].

For $\mu < c$, $\quad P_0(n,1) \;\sim\; e^{-n\mu/c}\,\dfrac{(n\,\mu/c)^n}{n!}, \quad \bar{J} \;\sim\; n\left[\dfrac{1}{\mu} \;-\; \dfrac{1}{c}\right] \;+\; \dfrac{n\,e^{-n\mu/c}}{\mu}\,\dfrac{(n\,\mu/c)^n}{n!}$
[from (63)].

3.5 THE EQUIVALENT RANDOM METHOD

This section describes Kosten's model and the so-called *equivalent random method*, based on it and due to Wilkinson [96].

In Kosten's model (see Figure 3), a Poisson stream offers a erlangs to a server group of n iid exponential servers, called the primary group. The overflow stream, called the O-stream, is offered to a second infinite server group of iid servers with the same service distribution as that of the first (finite) group, called the secondary group. The equilibrium offered load of the O stream is denoted by ℓ_O, and is measured in erlangs. The statistics of interest are the equilibrium time variance, σ_S^2, of the number of busy servers in the secondary group, S, and the peakedness, $z_{O,exp} = \sigma_S^2/\ell_O$.

It can be shown that the O-stream is a renewal traffic stream [90]. Consider now the continuous extension, $B(x,a)$, of the ordinary Erlang B function, $B(n,a)$ from (48), but with a continuous number of trunks, x, replacing the discrete number of trunks, n. When the O-stream is offered to the secondary server group, the resulting system is a $GI/M/\infty$ queue (*ibid.*)

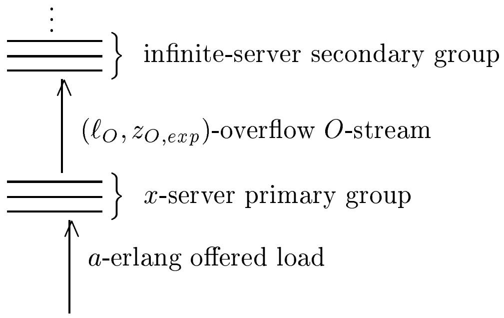

Figure 3: The Kosten Traffic Model.

The following fundamental formulae,

$$\ell_O = a\,B(x,a), \tag{76}$$

$$z_{O,exp}(\mu) = 1 - \ell_O + \frac{a}{x + 1 + \ell_O - a}, \tag{77}$$

are due to [56]. A direct derivation of these formulae proceeds as follows. Consider a $G/G/1/1$ queue, in which the arrival stream is a stationary simple point process (that is, arrivals occur singly). Let $N(t)$ be the number of arrivals in $(0,t]$, the origin being an arrival point, and let $M(t) = E[N(t)]$ be the attendant expectation function. The service distribution function is denoted by $B(t)$. Then the equilibrium probability, π, that an arrival finds the server busy is

$$\pi = \frac{\int_0^\infty M(t)\,dB(t)}{1 + \int_0^\infty M(t)\,dB(t)}.$$

To see that, note that in the $G/G/1/1$ queue, each busy cycle consists of just one arrival being served (the one inaugurating the busy cycle), with the remaining arrivals in that cycle being blocked; furthermore, since 0 is an arrival point, the server becomes busy at time 0. Hence, the numerator is the mean number of blocked arrivals during a service period, while the denominator is the total number of arrivals (served and blocked) during a busy cycle. The ratio is thus the requisite probability. In particular, for exponential service, one has $\pi = \mu\tilde{M}(\mu)/(1 + \mu\tilde{M}(\mu))$.

Consider a renewal stream with interarrival time density $a(t)$. Since, in this case, $\mu\,\tilde{M}(\mu) = (\tilde{a}(\mu)/(1 - \tilde{a}(\mu))$, one has

$$\pi \;=\; \tilde{a}(\mu). \tag{78}$$

Consider now the primary group in Kosten's model and let γ_j be the probability that the overflow of the jth server finds the $j + 1$th server busy. Then

$$\gamma_j \;=\; \frac{B(j+1, a)}{B(j, a)}. \tag{79}$$

The continuous extension of γ_j will be denoted by γ_x. Let $a(t)$ denote the equilibrium density of time between overflows in the O-stream. Then from (78) and (79),

$$\tilde{a}(\mu) \;=\; \frac{B(x+1, a)}{B(x, a)}. \tag{80}$$

Now, for any renewal traffic stream, A, one has [15]

$$\tilde{m}_A(\mu) \;=\; \frac{\tilde{a}(\mu)}{1 - \tilde{a}(\mu)}. \tag{81}$$

Suppose that O is a *renewal* overflow traffic stream in Kosten's setting. Substituting (81) into (38), and using the relations $\ell_O = \lambda_O/\mu$ (by definition) and $\ell_O = aB(x, a)$ from (76), yields

$$z_{O,exp}(\mu) \;=\; \frac{1}{1 - \tilde{a}(\mu)} \;-\; \ell_O. \tag{82}$$

Finally, combining (59) and (80) in (82) yields (77), with a modicum of algebra.

The Kosten formulae, (76) and (77), possess a unique solution for (x, a) in terms of ℓ_O and $z_{O,exp}(\mu)$ [52]. The following general traffic problem is of interest. A traffic stream, X, characterized by the offered load and peakedness parameters, $(\ell_X, z_{X,exp})$, is offered to a server group of c iid exponential servers, and the blocking probability is to be determined. This offered traffic is viewed as the overflow stream from a primary group of x servers resulting from a Poisson offered load of a erlangs, such that Eqs. (76) and (77) hold [96]. The corresponding parameters, (x, a), are called the *equivalent random parameters*; the "equivalence" is in the sense that the derived (x, a) are "equivalent" to the prescribed $(\ell_X, z_{X,exp})$, and the term "random" refers to a Poisson stream offered load. Figure 4 depicts the scheme.

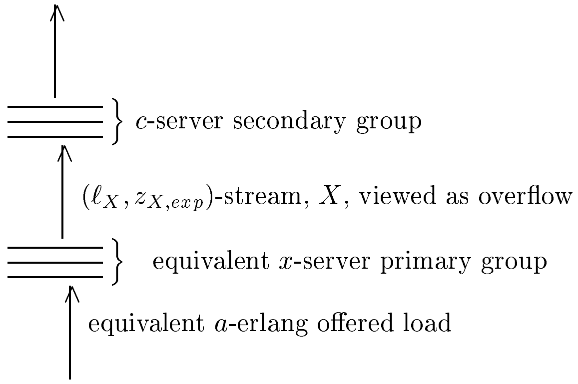

Figure 4: The Equivalent Random Scheme.

The requisite blocking probability, P_b, is given by

$$P_b \;=\; \frac{B(x+c, a)}{B(x, a)}. \tag{83}$$

Next, view (76) – (77) as a formal equation system in the equivalent parameters, (x, a), not necessarily in an O-stream context. To obtain (x, a), first solve for a in the equation

$$\ell_X \;=\; a\, B\left(a\, \frac{\ell_X + z_{X,exp}}{\ell_X + z_{X,exp} - 1} - \ell_X - 1, a \right), \tag{84}$$

and then compute x by substituting the obtained parameter, a, from (84) into

$$x \;=\; a\, \frac{\ell_X + z_{X,exp}}{\ell_X + z_{X,exp} - 1} \;-\; \ell_X \;-\; 1. \tag{85}$$

An initial approximation, (x_0, a_0), readily obtained from (84) and (85) by expansion in powers of $1/\ell_X$, is given by [81]

$$a_0 \;=\; \ell_X\, z_{X,exp} \;+\; 3\, z_{X,exp}\, (z_{X,exp} - 1).$$

An accurate computation of a can then be obtained from the recursion

$$\begin{aligned} a_1 \;&=\; a_0 \;-\; \frac{1}{B(x_0, a_0)}\, \frac{a_0\, B(x_0, a_0) - \ell_X}{x_0 + \ell_X + a - a_0 - (x_0 + \ell_X + 1)\, \ln \alpha_0}, \\ \alpha_0 \;&=\; \frac{x_0 + 1}{a_0} + B(x_0, a_0). \end{aligned}$$

When a stream characterized by $(\ell_X, z_{X,exp})$ is an O-stream, then x is an integer and the blocking probability P_b, given in (83), is exact. Otherwise, a non-integral value of x may be obtained. Nonetheless, (83) is used in practice as an approximation of the blocking probability.

When n independent streams, characterized by $(\ell_1, z_1), \ldots, (\ell_n, z_n)$, are superposed, then the superposition stream has parameters (ℓ, z) given by

$$\ell_X \;=\; \sum_{i=1}^{n} \ell_i, \quad z_{X,exp} \;=\; \frac{\sum_{i=1}^{n} \ell_i\, z_i}{\sum_{i=1}^{n} \ell_i}.$$

This follows on observing that each $\ell_i\, z_i$ is the variance of the number of busy servers produced by stream i when offered to an infinite server group of iid exponential servers with the same rate as that of the common server group, serving the superposition traffic stream.

If the target stream is the overflow of the common group, and is in turn offered (cascaded) to yet another group of c servers, then the peakedness of the target stream must be known, in order to compute its blocking probability. The requisite peakedness may be computed from (77) by first computing the equivalent random parameters, (x, a), and then considering the target stream to be the overflow from a group of $x + c$ servers with Poisson offered load of a erlangs.

Consider the time congestion statistic (the limiting fraction of time that all servers are busy), associated with a stream, characterized by $(\ell_X, z_{X,exp})$. The time congestion, P^*, in the cascaded system can be obtained, using the results in [11]. For any non-negative integer j, define $\delta_j(n, a)$ by the following recursion on n [49]

$$\delta_j(n,a) \;=\; \begin{cases} 1, & n = 0 \\ \dfrac{n\, \delta_j(n-1,a)}{a + j\, \delta_j(n-1,a)} + 1, & n > 0 \end{cases}$$

Letting $\delta_c(x, a)$ be the continuous extension of $\delta_j(n, a)$ for $j = c$ servers in the server group of the cascaded system, one has

$$P^* \;=\; \delta_c(x,a)\, B(x+c, a). \tag{86}$$

Of course, if the stream is not an O-stream, then (86) is taken as approximation. In order to compute $\delta_0(x, a)$ when x is not an integer, quadratic interpolation of $\ln \delta_c(x, a)$ is used. Recalling that $\lfloor x \rfloor$ denotes the integral part of x, and $\langle x \rangle$ the fractional part of x, one has

$$\delta_c(x,a) \;\simeq\; \delta_c(\lfloor x \rfloor, a)^{1-\langle x \rangle}\, \delta_c(\lfloor x \rfloor + 1, a)^{\langle x \rangle} \left(\frac{\delta_c(\lfloor x \rfloor + 1, a)^2}{\delta_c(\lfloor x \rfloor, a)\, \delta_c(\lfloor x \rfloor + 2, a)} \right)^{\langle x \rangle (1 - \langle x \rangle)/2}.$$

The quantities, $\delta_c(x, a)$, have been shown to obey the inequalities [49],

$$\delta_c(x,a) \;\geq\; \frac{1}{2} \left(1 \,+\, \frac{x-a+1}{c} \,+\, \sqrt{\left(1 + \frac{x-a+1}{c}\right)^2 + 4\, \frac{c\,(a-1) - x}{c^2}} \right),$$

$$\delta_c(x,a) \;\leq\; \frac{1}{2} \left(1 \,+\, \frac{x-a}{c} \,+\, \sqrt{\left(1 + \frac{x-a}{c}\right)^2 + \frac{4\,a}{c}} \right),$$

which are useful for real-time computations. In particular, the lower bound is often an excellent approximation.

3.6 THE HAYWARD APPROXIMATION

The equivalent random method is subject to the limitation that the service distribution is assumed to be exponential. This limitation may be removed by use of the so-called Hayward approximation of blocking probabilities for general service distributions.

Consider a c server system of iid servers with service distribution $F(x)$ and rate μ. A traffic stream, X, with parameters (ℓ_X, z_{X,F_0}) is offered to the server group. The *Hayward approximation* for the blocking probability, P_b, is given by [32]

$$P_b \;\simeq\; B\left(\frac{c}{z_{X,F_0}(\mu)}, \frac{a}{z_{X,F_0}(\mu)}\right), \tag{87}$$

where the function $B(\cdot,\cdot)$ above is the Erlang B function from (48).

Measurements of the peakedness functional are often made on a server whose service distribution, say G, is not in the requisite family, F_0. Thus, it is important to transform the measured information, that is, to transform z_{X,G_0} into z_{X,F_0}. To this end one employs the change of service distribution theory surveyed in Section 3.1.3 in the context of Hayward's approximation, (87). As in the equivalent random procedure, it is important to calculate the peakedness, z_{Q,F_0}, of the overflow traffic stream, Q; note, though, that Q is generally not an O-stream. A modification of (77) yields [32]

$$z_{Q,F_0}(\mu) \;=\; z_{X,F_0}(\mu)\left[1 \;-\; \ell_1 \;+\; \frac{a_1}{c_1+1+\ell_1+a_1}\right], \tag{88}$$

where $c_1 = c/z_{X,F_0}(\mu)$, $a_1 = a/z_{X,F_0}(\mu)$ and $\ell_1 = a_1\, B(c_1, a_1)$. The peakedness in (88) can be further transformed to other service distributions, as described in Section 3.1.3, and the results used to approximate the blocking probability (87).

3.7 WAITING TIMES FOR GENERAL TRAFFIC

A traffic process with an interarrival time sequence, $A = \{A_n\}$, is termed a *general traffic (GT) stream*, if the only known parameters are the offered load, ℓ_A, and the peakedness, $z_{A,exp}$, with respect to exponential service distributions. Consider a GT stream with a stationary interarrival time sequence, A, of rate λ_A. Suppose that the GT stream is offered to a finite-server group of k iid exponential servers with service rate μ, and that the waiting room is infinite. This queueing model will be denoted by $GT/M/k/\infty$. The objective here is to calculate or approximate the equilibrium distribution, F_W, of the waiting time in queue (excluding service), W.

We begin with renewal traffic for two reasons: First, F_W can then be computed exactly, and second, the method used provides the basis for approximating F_W in non-renewal cases. Accordingly, we start with the $GI/M/k/\infty$ queue. Let $r \in (0,1)$ be the root of

$$[1-r]\, z_{A,exp}(\mu k(1-r)) \;=\; 1 \;-\; \frac{\ell_A}{k}. \tag{89}$$

Additionally, the Laplace transform of the renewal density function of A is

$$\tilde{m}_A(s) \;=\; z_{A,exp}(s) \;-\; 1 \;+\; \frac{\lambda_A}{s}, \tag{90}$$

and the transform of the interarrival time density function, $a(t)$, is

$$\tilde{a}(s) \;=\; 1 \;-\; \frac{1}{z_{A,exp}(s) + \lambda_A/s}. \tag{91}$$

From [90], one has

$$F_W^c(t) \;=\; \frac{D}{1-r}\, e^{-\mu k(1-r)t}, \tag{92}$$

$$E[W] \;=\; \frac{D}{\mu\, k\,(1-r)^2}, \tag{93}$$

where $D^{-1} = \dfrac{1}{1-r} + \displaystyle\sum_{j=1}^{k} \frac{\binom{k}{j}}{M_j\,(1-\tilde{a}(j\mu))}\,\frac{k\,(1-\tilde{a}(j\mu))-j}{k\,(1-r)-j}$, and $M_j = \displaystyle\prod_{i=1}^{j} \tilde{m}_A(i\mu)$.

A significant aspect of (89) through (93) is that the $GI/M/k/\infty$ queue is analyzed using only ℓ_A and $z_{A,exp}$. Thus, the same analysis could be formally carried out for the $GT/M/k/\infty$ queue, by assuming that F_W may be approximated by an exponential distribution of the form (92), which is exact for the $GI/M/k/\infty$ queue [26]. The resulting F_W^c from (92) and $E[W]$ from (93) will then serve as approximations.

If only a single value of $z_{A,exp}(\mu)$ is known, for example at $\mu = \mu_1$, then the *three moment match* technique [49] may be used to construct an *interrupted Poisson process* (IPP) [57], which in turn provides a usable choice of $z_{A,exp}(\mu)$ with a prescribed value at μ_1. This may now be used in (89)–(93) to carry out an approximate analysis of the $GT/M/k/\infty$ model.

If at least two values of $z_{A,exp}(\mu)$ are known, then one may assume that the expectation density function is of the form

$$m_A(t) \;=\; \lambda_A \;+\; C\,e^{-\alpha t},$$

for some positive constants, C and α, so that for any stationary orderly traffic stream, A,

$$z_{A,exp}(\mu) \;=\; 1 \;+\; \frac{C}{\alpha+\mu}.$$

Thus practical means are available for estimating queueing delays in a finite-server group of exponential servers, based only on the traffic parameters $(\ell_A, z_{A,exp})$.

3.8 FREQUENCY DOMAIN APPROACH TO TRAFFIC

The Frequency Domain Approach (FDA) focuses on second-order statistics of offered traffic and their effect on queue response to that traffic [61, 62, 86]; it has been motivated by the need to characterize multimedia traffic in high-speed networks. FDA is distinguished by the fact that it directly utilizes the frequency domain (the traffic spectral functions), and advocates their use as a unified traffic measurement for analyzing and controlling queueing systems with heterogeneous offered traffic.

Analogously to periodic input functions in signal processing, elements of constant, sinusoidal, rectangular pulse, triangle pulse, and their superpositions are used in Li and Hwang [61] to represent various second-order traffic statistics, and to observe their effect on queueing performance. The main finding is that only the low frequencies in the traffic spectrum have a significant effect on queueing statistics. However, this approach has limited applications, since it does not capture the stochastic aspects of a prescribed traffic process. To this end, multi-state MMPP traffic are used in Sheng and Li [86] to characterize binary sources (on/off traffic), and to study the effect of their second-order statistical properties on queue length and loss rate statistics.

A modeling technique for constructing Markovian traffic processes, which match a prescribed power spectrum, is introduced in Li and Hwang [62]. Let Q be the infinitesimal generator (transition rate matrix) of an N-state Markov chain, which modulates the Poisson rate of the traffic process with rate vector $\vec{\gamma}$. It is shown that the eigenstructure of Q characterizes the effect of such traffic on queueing performance. Assume that Q is diagonalizable with eigenvalues $\vec{\lambda} = (\lambda_0, \lambda_1, ..., \lambda_{N-1})$. The power spectrum of the random (Poisson) rate process can then be expressed by

$$P(\omega) \;=\; \sum_{j=0}^{N-1} \frac{-2\,\psi_j\,\lambda_j}{\lambda_j^2 + \omega^2}. \tag{94}$$

Essentially, each eigenvalue, λ_j, contributes a bell-shaped component in the traffic rate power spectrum. Each bell is described by the average power, ψ_j, the central frequency, $Im[\lambda_j]$, and the half-power bandwidth, $-2\,Re[\lambda_j]$. Denoting $\vec{\psi} = (\psi_0, \psi_1, ..., \psi_{N-1})$, the traffic power spectrum is characterized by the pair, $(\vec{\psi}, \vec{\lambda})$, and the goal of FDA is to construct $(Q, \vec{\gamma})$ from $(\vec{\psi}, \vec{\lambda})$ — a rather difficult task, involving the so-called inverse eigenvalue problem. The technique in Li and Hwang [62] uses a special class of Markov chains, called *circulants*, with infinitesimal generators of the form

$$Q \;=\; \begin{pmatrix} a_0 & a_1 & a_2 & \cdots & a_{N-1} \\ a_{N-1} & a_0 & a_1 & \cdots & a_{N-2} \\ a_{N-2} & a_{N-1} & a_0 & \cdots & a_{N-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_1 & a_2 & a_3 & \cdots & a_0 \end{pmatrix} \tag{95}$$

where each row is a circular right-shift of the previous one. A circulant-type modulated traffic rate process is then characterized by the pair, $(\vec{a}, \vec{\gamma})$, $\vec{a}$ being the first row of Q. Its power spectrum was shown to be

$$\vec{\lambda} \;=\; \sqrt{N}\,\vec{a}\,F^{\star}, \quad \vec{\psi} \;=\; \frac{1}{N}\,|\vec{\gamma}\,F^{\star}|^2, \tag{96}$$

where F is a Fourier matrix with $F_{k,j} = \frac{1}{\sqrt{N}}\,e^{-\frac{2\pi k j}{N}\,i}$, and $i = \sqrt{-1}$. Then $F^{-1} = F^{\star}$, where $F^{\star}$ is the conjugate transpose of F. More importantly, one can deduce the corresponding pair, $(\vec{a}, \vec{\lambda})$, that gives rise to a prescribed power spectrum, characterized by a pair, $(\vec{\lambda}, \vec{\psi})$, by taking the following constrained inverse discrete Fourier transform,

$$\begin{aligned} \vec{a} \;&=\; \frac{1}{\sqrt{N}}\,\vec{\lambda}\,F \quad \text{subject to } a_j \ge 0 \text{ for } j > 0 \text{ and } a_0 = -\textstyle\sum_{j=1}^{N-1} a_j, \\ \vec{\gamma} \;&=\; \sqrt{N}\,\vec{\beta}\,F \quad \text{subject to } \gamma_j \ge 0,\ \forall j, \text{ with } \beta_j = \sqrt{\psi_j}\,e^{i\theta_j}, \end{aligned} \tag{97}$$

where $\vec{\theta} = (\theta_0, \theta_1, ..., \theta_{N-1})$ is a phase vector. Hence, $\vec{a}$ depends only on $\vec{\lambda}$, and $\vec{\gamma}$ only on $(\vec{\psi}, \vec{\theta})$. It should be noted that while $\vec{\theta}$ is unrelated to the power spectrum, it influences the steady-state distribution and higher-order traffic statistics [62]. The numerical analyses in [62, 85] are based on the Folding Algorithm, which is a fast computational method for analysis of finite quasi-birth-death process with level-dependent transitions [98].

A significant advantage of using circulants in FDA is the ability to identify the impact of first-order (marginal distribution), second-order (power spectrum or autocorrelation function), and higher-order traffic statistics (such as the bispectrum and trispectrum) on queue length and loss rate statistics. Interestingly, only second-order traffic statistics appear to have a significant effect on queueing performance, with the low frequencies exerting the most effect. This observation has important potential applications to traffic measurement, especially since first-order and second-order statistics are much easier to measure than higher-order ones. FDA has been applied, in this vein, to traffic rate control [63], to link capacity allocation and network control [64], and to the design and analysis of buffer congestion control [85].

3.9 EFFECTIVE BANDWIDTH OF TRAFFIC

Most communications services are subject to performance constraints designed to guarantee a minimal quality of service (QOS). For example, the delivery of a video frame to a monitor must be both timely and not overly lossy. Consider a general traffic stream offered to a deterministic server, and assume that some prescribed parametrized performance constraints are required to hold. The *effective bandwidth* of the traffic stream corresponds to the minimal deterministic service rate, required to meet these constraints. Queueing-oriented performance constraints include bounds on such statistics as queueing delay quantiles or averages, server utilization, overflow probabilities, etc. [53]. This approach has the advantage of focusing directly on the relevant performance criteria rather than on the statistics of the offered traffic.

To illustrate this viewpoint, suppose that a traffic stream shares a link with buffering capacity B and bandwidth c. Suppose further that the QOS constraint is that the overflow probability satisfy

$$P\{W > B\} \leq e^{-\delta B}, \tag{98}$$

where W denotes the stationary workload in the buffer and δ is a QOS parameter. In this example, the corresponding effective bandwidth of the traffic stream, $\alpha(\delta)$, is the minimal deterministic service guaranteeing (98). Operationally, if the goal is to ensure that the overflow probability incurred by a traffic stream offered to a deterministic server is small (say, 10^{-9} for a particular δ), then a link bandwidth of at least $\alpha(\delta)$ should be allocated to carry that traffic. As $\delta \geq 0$ increases, the overflow constraint becomes more stringent and the effective bandwidth increases from the mean arrival rate to the peak rate. Viewed this way, the effective bandwidth concept serves as a compromise between two alternative bandwidth allocation schemes, representing a pessimistic and a optimistic outlook. The strict one allocates bandwidth based on the stream peak rate, seeking to eliminate losses, whereas the lenient one allocates bandwidth based on the stream average rate, merely seeking to guarantee stability.

Generally, effective bandwidths are difficult to compute, as their value depends both on traffic stream statistics as well as possible interactions with cross traffic traversing the same system. Occasionally, a direct calculation of the effective bandwidth is feasible; a case in point are Markov fluid sources, where the overflow probability can be computed explicitly as a function of the service rate via spectral expansions [39, 35, 27].

An alternative approach is to consider asymptotically large buffers B and find approximate effective bandwidths for the streams, see [12, 54, 19, 95]. For example, consider an unbounded buffer shared by a heterogeneous mix of multiplexed streams, of which n_j are of type $j \in J$, where J is the set of traffic types. For service rate c, it can be shown that

$$\sum_{j \in J} n_j \alpha_j(\delta) \;\le\; c \iff \lim_{B \to \infty} \frac{1}{B} \log P\{W > B\} \;\le\; -\delta,$$

where $\alpha_j(\delta)$ is the *effective bandwidth* of stream j and depends only on the statistics of stream j, say, in terms of the interarrival times, $\{A_i^j\}_{i=1}^{\infty}$. This approximation can be established for a relatively large class of arrival streams which are stationary, ergodic and have weak dependencies such as Markov or mixing processes. For example, a traffic stream with slotted arrivals has an effective bandwidth, given by

$$\alpha_j(\delta) \;=\; \frac{\Lambda_j(\delta)}{\delta}, \quad \text{where} \quad \Lambda_j(\delta) \;=\; \lim_{n \to \infty} \frac{1}{n} \log E \exp[\delta \sum_{i=1}^{n} A_i^j],$$

which can be computed for many standard traffic models, including Markov fluids and MMPP.

The notion of effective bandwidth provides a useful tool for studying resource requirements of telecommunications services and the impact of different management schemes on network performance. A study of the impact of filtering, traffic discard, and traffic shaping on the effective bandwidth of traffic streams can be found in [18]. Extensions of these results to networks of queues are proposed in [19].

4 DISCUSSION AND OPEN PROBLEMS

From a practical point of view, stochastic models of traffic streams are considered relevant to network traffic engineering and performance analysis, to the extent that they are able to predict system performance measures to a reasonable degree of accuracy; additionally, a practitioner's confidence in a traffic model is greatly enhanced, if the model can capture visual features of the empirical traffic, in addition to approximating its statistics. The fundamental systems, of which traffic is a key ingredient, are queueing systems. While traditional analytical models of traffic were often devised and selected for the analytical tractability they induced in the corresponding queueing systems, this selection criterion is largely absent from recent traffic models. In particular, queueing systems with offered traffic consisting of autoregressive-type processes (Section 2.4), TES processes (Section 2.5), and self-similar processes (Section 2.6) are difficult to solve analytically. Consequently, these are only used to drive Monte Carlo simulation models. On the other hand, some fluid models (Section 2.3) are analytically tractable, but only subject to considerable restrictions. Similarly, FDA models (Section 3.8) and effective bandwidth models (Section 3.9) are largely analytically tractable, but at the cost of simple models, and restrictive modeling assumptions. Thus, the most significant traffic research problem is to solve analytically induced queueing systems; or in the absence of a satisfactory solution, to devise approximate traffic models which lead to analytically tractable queueing systems.

Traditional traffic models (Sections 2.1 – 2.4) have served admirably in advancing traffic engineering and understanding performance issues, primarily in traditional telephony.

The advent of modern high-speed communications networks (e.g., B-ISDN under ATM) is ushering in a dramatically heterogeneous traffic mix. The inherent burstiness of several important B-ISDN services (mainly compressed video and file transfer) is bringing to the fore some serious modeling inadequacies of traditional models, particularly in regard to temporal dependence, both short-range and long-range. This realization has brought about renewed interest in traffic modeling and has driven the development of new models utilizing non-traditional paradigms, such as TES processes and self-similar processes.

The importance of empirical second-order traffic statistics, such as indices of dispersion (Section 3.1.2) and peakedness (Section 3.1.3), has already been recognized in traditional traffic engineering, and again in recent work on traffic models, including TES, self-similar and FDA models. Of those, only the TES modeling paradigm addresses itself directly to capturing both the empirical marginal distribution (histogram) and empirical autocorrelation function, simultaneously, and additionally calls for qualitative "resemblance" of sample paths. The availability of TES modeling software, incorporating both heuristic and algorithmic fitting procedures, render the TES methodology a practical paradigm for constructing traffic models directly from empirical data, but the absence of corresponding analytically tractable queueing models currently restricts it to Monte Carlo simulations. An important traffic research problem, motivated by the TES approach, is to find when first-order and second-order statistics are insufficient for traffic characterization, and whether using higher-order spectral functions (such as the bispectrum and trispectrum) can remedy such characterization problems.

Initially, the validity and efficacy of traffic models for high-speed networks, was difficult to assess due to the paucity or unavailability of empirical data, but more recently, increasing volumes of traffic measurements from working high-speed networks (e.g., CCS/SS7 at 56 kbps, ISDN at 1.5 Mbps, Ethernet at 10 Mbps) have been made available to researchers. In particular, the traffic studies mentioned in Section 2.6 suggest that certain Ethernet traffic has distinguishing features, often referred to as *fractal properties*; these typically pertain to notions such as slowly decaying variances, long-range dependence, $1/f$-noise or self-similarity. Statisticians are now aware that ignoring long-range dependence can have drastic consequences for many statistical methods [6]. However, traffic engineers and network managers will only be convinced of the practical relevance of fractal traffic models by direct arguments, concerning the impact of fractal properties on network performance. Thus, fractal traffic (including stochastic modeling, statistical inference, synthetic traffic generation and queueing analysis) opens a new and challenging area of mathematical research, where few results are available, and what little is known (e.g., [33, 59]) is mostly preliminary in nature and illustrates the difficulties presented by these models.

Recent results reported in [23, 79] demonstrate that performance measures of queueing systems with fractal traffic can differ drastically from those predicted by corresponding systems with traditional traffic models (see also [30, 59]). As in the case of bursty TES traffic [66] traditional models are way too optimistic as compared to self-similar models. A case in point is the tail behavior of the steady-state queue-length distribution $P\{Q > x\}$ (x large) in a single-server, infinite-capacity queue [79, 23]. For Markovian offered traffic, these tail probabilities are known to be *asymptotically exponential* [1, 36], i.e.,

$$P\{Q > x\} \sim e^{-\eta x} , \text{ as } x \to \infty, \tag{99}$$

where $\eta > 0$ is called the *asymptotic decay rate.* Recall that relation (99) underlies the concept of effective bandwidth, where admision control or service capacity allocation are based on tail probabilities of select random variables (e.g., Eq. (98)). In contrast to (99), traffic streams with long-range dependence (in particular, fractional Brownian motion-based models) give rise to corresponding *Weibull-like* asymptotic tail probabilities, i.e.,

$$P\{Q > x\} \sim e^{-\gamma x^{\beta}}, \text{ as } x \to \infty, \tag{100}$$

where γ is a constant, and $\beta = 2 - 2H \in (0, 1)$ [79, 23]. Eqs. (99) and (100) have very different asymptotic characteristics, the former providing relatively optimistic predictions as compared to the latter; see also [33, 79] for additional queueing examples. An open problem of considerable interest is whether other traffic models tend to give rise to conservative performance measures, as compared to their empirical counterparts.

If second-order or fractal properties (e.g., the Joseph Effect and Noah Effect) lead to fundamentally distinct system behavior, then the ability of tractable traffic models to capture such properties and their impact (say, on queueing statistics) becomes of interest. For example, while non-fractal models (such as Markov-based ones) have inherently short-range dependence, it is nevertheless known that adding parameters can lead to models with "approximate" fractal features, e.g., Markovian models with "supplemental" states, and TES-based or FDA-based models with low frequency components of "enhanced" magnitude, particularly the DC component. A judicious choice of a traffic model could lead to tractable queueing models capable of approximating their intractable counterparts. However, one can expect to give up model parsimony, and this approach may work for some performance aspects and fail for others.

Overall, the variety of traffic classes introduced by emerging high-speed communications networks will doubtless prove to be a rich source of research, both theoretical and applied. One can fully expect traffic engineers to call upon statisticians and probabilists to develop new theoretical and applied tools to assist in traffic handling at every stage of communications systems life cycle, including network design, analysis and operation.

Acknowledgments

We are grateful to San-Qi Li for the material on the frequency domain approach, and to Gustavo de Veciana for the material on effective bandwidths. We also thank Robert Cooper, Gustavo de Veciana and Robert Maier for reading and commenting on the manuscript. Special thanks are due to Tomasz Rolski for a careful reading of the paper and for numerous useful suggestions.

References

[1] Abate, J., Choudhury, G.L. and Whitt, W., “Asymptotics for Steady-State Tail Probabilities in Structured Markov Chains”, *Stochastic Models* **10** (1994), 99–143.

[2] Anick, D., Mitra, D. and Sondhi, M.M., “Stochastic Theory of a Data-Handling System with Multiple Sources”, *The Bell System Technical Journal* **61:8** (1982), 1871–1894.

[3] Asmussen, S. *Applied Probability and Queues*, Wiley, 1987.

[4] Asmussen, S. and Koole, G. “Marked Point Processes as Limits of Markovian Arrival Streams”, *J. Appl. Prob.* **30** (1993), 365–372.

[5] Bendat, J.S. and Piersol, A.G., *Random Data*, Wiley, 1986.

[6] Beran, J., “Statistical Methods for Data with Long-Range Dependence”, *Statistical Science* **7:4** (1992), 404–427.

[7] Beran, J., Sherman, R., Taqqu, M.S. and Willinger, W., “Long-Range Dependence in Variable-Bit-Rate Video Traffic”, *IEEE Trans. on Communications* **43** (1995), 1566–1579.

[8] Błaszczyszyn, B., Rolski, T. and Schmidt, V., “Light-Traffic Approximations in Queues and Related Stochastic Models”, Chapter 15 in *Advances in Queueing: Theory, Methods and Open Problems* (J.H. Dshalalow, Ed.), 379–406, CRC Press, 1995.

[9] Box, G.E.P., and Jenkins, G.M., *Time Series Analysis: Forecasting and Control*, Prentice Hall, 1976.

[10] Brémaud, P., *Point Processes and Queues: Martingale Dynamics*, Springer-Verlag, 1981.

[11] Brockmeyer, E., “The Simple Overflow Problem in the Theory of Telephone Traffic”, *Teleteknik* **5** (1954), 361–74.

[12] Chang, C.S., “Stability, Queue Length and Delay, Part II: Stochastic Queueing Networks” IBM Research Report No. RC 17709, T.J. Watson Research Center, Yorktown Heights, New York, 1992.

[13] Cinlar, E., *Introduction to Stochastic Processes*. Prentice-Hall, 1975.

[14] Cooper, R.B., *Introduction to Queueing Theory*, North Holland, New York, 1981.

[15] Cox, D.R., *Renewal Theory*, Methuen and Co., London, 1962.

[16] Cox, D.R., ”Long-Range Dependence: A Review”, in *Statistics: An Appraisal*, H.A. David and H.T. David (Eds.), The Iowa State University Press, Ames, Iowa, 1984, 55–74.

[17] Davis, P.J. and Rabinowitz, P., *Methods of Numerical Integration*, Academic Press, New York, 1975.

[18] de Veciana, G. and Walrand, J., "Effective Bandwidths: Call Admission, Traffic Policing and Filtering for ATM Networks", accepted to *Queueing Systems* (1994).

[19] de Veciana, G., Courcoubetis, C. and Walrand, J., "Decoupling Bandwidths for Networks: A Decomposition Approach to Resource Management for Networks", Memorandum UCB/ERL M93/50, U.C. Berkeley, 1993.

[20] Dellacherie, C. and Meyer, P.-A., *Probabilities and Potential B*, North-Holland, 1982.

[21] Devroye, L., *Non-Uniform Random Variate Generation.* Springer-Verlag, 1986.

[22] Doetsch, G., *Theorie und Anwendung der Laplace-Transformation*, Dover Publications, New York, 1943.

[23] Duffield, N.G. and O'Connell, N., "Large Deviations and Overflow Probabilities for the General Single-Server Queue, with Applications", Dublin Institute for Advanced Studies, preprint, 1993.

[24] Duffy, D.E., McIntosh, A.A., Rosenstein, M. and Willinger, W., "Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks", *IEEE J. Select Areas Communications* **12:3** (1994), 544–551.

[25] Eckberg, A.E., "Generalized Peakedness of Teletraffic Processes", *Proc. 10-th International Teletraffic Congress*, Montreal, Canada, 1983.

[26] Eckberg, A.E., "Approximations for Bursty (and Smoothed) Arrival Delays based on Generalized Peakedness," *Proc. 11-th International Teletraffic Congress*, Kyoto, Japan, 1985.

[27] Elwalid, A.I. and Mitra, D., "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks", *IEEE/ACM Trans. Networking* **1:3** (1993), 329–343.

[28] Fendick, K. and Whitt, W., "Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue", *Proc. of the IEEE* **77** (1989), 171–194.

[29] Fendick, K.W., Saksena, V.R. and Whitt, W., "Dependence in Packet Queues", *IEEE Trans. on Comm.* **37** (1989), 1173–1183.

[30] Fowler, H.J. and Leland, W.E., "Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management" *IEEE J. Select. Areas Commun.* **SAC-9:7** (1991), 1139–1149.

[31] Franken, P., Koenig, D., Arndt, U. and Schmidt, V., *Queues and Point Processes*, Akademie-Verlag, 1981.

[32] Fredericks, A.A., "Congestion in Blocking Systems – A Simple Approximation Technique," *Bell System Technical Journal* **59:6** (1980), 805–827.

[33] Garrett, M.W. and Willinger, W., "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic", *Proc. of the ACM Sigcomm '94*, London, UK, 1994, 269–280.

[34] Geist, D. and Melamed, B., "TEStool: An Environment for Visual Interactive Modeling of Autocorrelated Traffic", *Proceedings of the 1992 International Conference on Communications*, Chicago, Illinois, 1992, 1285–1289.

[35] Gibbens, R.J. and Hunt, P.J., "Effective Bandwidths for the Multi-type UAS Channel", *Queueing Systems* **9** (1991), 17–28.

[36] Glynn, P.W. and Whitt, W., "Logarithmic Asymptotics for Steady-State Tail Probabilities in a Single-Server Queue", in *Studies in Applied Probability, Essays in honour of Lajos Takacs*, J. Galambos and J. Gani (eds.), Applied Probability Trust, Sheffield, England, 1994, 131–156. Also in *J. Appl. Prob.* special volume **31A**, 1994.

[37] Granger, C.W.J. and Joyeux, R., "An Introduction to Long-Memory Time Series Models and Fractional Differencing", *Time Series Anal.* **1** (1980), 15–29.

[38] Granger, C.W.J., "Long Memory Relationships and the Aggregation of Dynamic Models", *J. Econometrics* **14** (1980), 227–238.

[39] Guérin, R., Ahmadi, H. and Naghshineh, M., "Equivalent Capcity and its Application to Bandwidth Allocation in High-Speed Networks", *IEEE JSAC* **9** (1991), 968–981.

[40] Gusella, R., "Characterizing the Variability of Arrival Processes With Indexes of Dispersion", *IEEE J. on Selected Areas in Communications*, **9:2** (1991), 203–211.

[41] Hardy, G.H., Littlewood, J.E. and Polya, G., *Inequalities*, Cambridge University Press, 1959.

[42] Heffes, H. and Lucantoni, D.M., "A Markov Modulated Characterization of Packetized Voice and Data and Related Statistical Multiplexer Performance", *IEEE J. on Selected Areas in Communications* **SAC-4** (1986), 856–868.

[43] Heyman, D., Tabatabai, A. and Lakshman, T.V., "Statistical analysis and simulation study of video teletraffic in ATM networks", *IEEE Trans. Circuits and Systems for Video Technology* **2** (1992), 49–59.

[44] Hosking, J.R.M., "Fractional Differencing", *Biometrika* **68** (1981), 165–176.

[45] Hurst, H. E., "Long-Term Storage Capacity of Reservoirs", *Trans. Amer. Soc. Civil Engineers* **116** (1951), 770–799.

[46] Jacod, J., *Calcul Stochastique et Problème de Martingales*, Lecture Notes in Math. 714, Springer-Verlag, 1974.

[47] Jagerman, D.L., "Some Properties of the Erlang Loss Function," *Bell System Technical Journal* **53:3** (1974), 525–551.

[48] Jagerman, D.L., "Methods in Traffic Calculations", *Bell System Technical Journal* **63:7** (1984), 1283–1310.

[49] Jagerman, D.L., "Laplace Transform Inequalities with Application to Queueing", *Bell System Technical Journal* **64:7** (1985), 1755–1764.

[50] Jagerman, D. L. and Melamed, B., "The Transition and Autocorrelation structure of TES Processes Part I: General Theory", *Stochastic Models* **8:2** (1992), 193–219.

[51] Jagerman, D. L. and Melamed, B., "The Transition and Autocorrelation structure of TES Processes Part II: Special Cases", *Stochastic Models* **8:3** (1992), 499–527.

[52] Jagers, A.A. and Van Doorn, E.A., "On the Continued Erlang Loss Function," *Operations Research Letters* **5:1** (1986), 43–46.

[53] Kelly, F.P., "Effective Bandwidths at Multi-Class Queues", *Queueing Systems* **9** (1991), 5–16.

[54] Kesidis, G., Walrand, J. and Chang, C.S., "Effective Bandwidths for multiclass Markov Fluids and Other ATM Sources", *IEEE/ACM Trans. Networking.* **1:4** (1993), 424–428.

[55] Kobayashi, H. and Ren, Q., "A Mathematical Theory for Transient Analysis of Communications Networks", *IEICE Transactions on Communications* **E75-B:12** (1992), 1266–1276.

[56] Kosten, L., *Stochastic Theory of Service Systems*, Pergammon Press, 1973.

[57] Kuczura, A., "The Interrupted Poisson Process as an Overflow Process," *Bell System Technical Journal* **52:3** (1973), 437–448.

[58] Lee, D.S., Melamed, B., Reibman, A. and Sengupta, B., "TES Modeling for Analysis of a Video Multiplexor", *Performance Evaluation* **16** (1992), 21–34.

[59] Leland, W.E., Taqqu, M.S., Willinger, W. and Wilson, D.V., "On the Self-Similar Nature of Ethernet Traffic (Extended Version)", *IEEE/ACM Trans. on Networking* **2:1** (1994), 1–15.

[60] Levy, J.B. and Taqqu, M.S., "On Renewal Processes Having Stable Inter-Renewal Intervals and Stable Rewards", *Ann. Sc. Math. Quebec* **11** (1987), 95–110.

[61] Li, S.Q. and Hwang, C.L., "Queue Response to Input Correlation Functions: Discrete Spectral Analysis," *IEEE/ACM Trans. on Networking* **1:5** (1993), 522–533.

[62] Li, S.Q. and Hwang, C.L., "Queue Response to Input Correlation Functions: Continuous Spectral Analysis," *IEEE/ACM Trans. on Networking* **1:6** (1993), 678–692.

[63] Li, S.Q. and Chong, S., "Fundamental Limits of Input Rate Control in High-Speed Network," *Proc. IEEE Infocom '93*, April 1993, San Fracisco, CA, 662–671.

[64] Li, S.Q., Chong, S., Hwang, C.L. and Zhao, X., "Link Capacity Allocation and Network Control by Filtered Input Rate in High Speed Networks," *Proc. IEEE Globecom '93*, December 1993, Houston, Texas, 744–550.

[65] Lighthill, M.J., *Introduction to Fourier Analysis and Generalised Functions*, Cambridge University Press, 1960.

[66] Livny, M., Melamed, B. and Tsiolis, A.K., "The Impact of Autocorrelation on Queuing Systems", *Management Science* **39** (1993), 322–339.

[67] Lucantoni, D.M., Meier-Hellstern, K.S. and Neuts, M.F., "A Single-Server Queue with Server Vacations and a Class of Non-Renewal Arrival Processes", *Adv. Appl. Prob.* **22** (1990), 676–705.

[68] Larson, H.J. and Shubert, B.O., *Probabilistic Models in Engineering Sciences*, John Wiley, 1979.

[69] Mandelbrot, B.B., "Self-Similar Error Clusters in Communication Systems and the Concept of Conditional Stationarity", *IEEE Trans. Communications Technology* **COM-13** (1965), 71–90.

[70] Mandelbrot, B.B. and Van Ness, J.W., "Fractional Brownian Motions, Fractional Noises and Applications", *SIAM Review* **10** (1968), 422–437.

[71] Mandelbrot, B.B., "Long-Run Linearity, Locally Gaussian Processes, H-Spectra and Infinite Variances", *Intern. Econom. Review* **10** (1969), 82–113.

[72] Mandelbrot, B.B. and Wallis, J.R., "Some Long-Run Properties of Geophysical Records", *Water Resources Research* **5** (1969), 321–340.

[73] Mandelbrot, B.B. and Taqqu, M.S., "Robust R/S Analysis of Long Run Serial Correlation", *Proc. 42nd Session ISI, vol. XLVIII, Book 2* (1979), 69–99.

[74] Mandelbrot, B.B., *The Fractal Geometry of Nature*. Freeman, New York, 1983.

[75] Meier-Hellstern, K., Wirth, P.E., Yan, Y.-L. and Hoeflin, D.A., "Traffic Models for ISDN Data Users: Office Automation Application", *Proc. 13-th International Teletraffic Congress*, Copenhagen, Denmark, 1991, 167–172.

[76] Melamed, B. (1993) "An Overview of TES Processes and Modeling Methodology", in *Performance Evaluation of Computer and Communications Systems*, (L. Donatiello and R. Nelson, Eds.), 359–393, Springer-Verlag Lecture Notes in Computer Science.

[77] Melamed, B. and Sengupta, B., "TES Modeling of Video Traffic", *IEICE Transactions on Communications* **E75-B:12** (1992), 1292–1300.

[78] Melamed, B. and Yao, D.D., "The ASTA Property", Chapter 7 in *Advances in Queueing: Theory, Methods and Open Problems* (J.H. Dshalalow, Ed.), 195–224, CRC Press, 1995.

[79] Norros, I., "A Storage Model with Self-Similar Input", *Queueing Systems* **16** (1994), 387–396.

[80] Ramamurthy, G. and Sengupta, B., "Modeling and Analysis of a Variable Bit Rate Video Multiplexor", *Proceedings of INFOCOM '92*, Florence, Italy, 1992, 817–827.

[81] Rapp, L.Y., "Planning of Junction Networks in a Multi-Exchange Area – Part I", *Ericson Technics* **20** (1962), 22–130.

[82] Reibman, A.R., "DCT-based Embedded Coding for Packet Video", *Signal Processing: Image Communication* **3** (1991), 231–237.

[83] Schassberger, R.A. "Investigation of Queuing and Related Systems with the Phase Method", *Proc. 7-th International Teletraffic Congress*, 1973.

[84] Sen, P., Maglaris, B., Rikli, N.-E. and Anastassiou, D., "Models for Packet Switching of Variable-Bit-Rate Video Sources", *IEEE J. on Selected Areas in Communications* **7:5** (1989), 865–869.

[85] Sheng, H.D. and Li, S.Q., "Spectral Analysis of Packet Loss Rate at Statistic Multiplexer for Multimedia Services," *IEEE/ACM Trans. on Networking* **2:1** (1994), 53–65.

[86] Sheng, H.D. and Li, S.Q., "Second Order Effect of Binary Sources on Characteristics of Queue and Loss Rate," *IEEE Trans. on Communications* **42:3** (1994), 1162–1173.

[87] Sigman, K. *Stationary Marked Point Processes*, Chapman and Hall, 1994.

[88] Sriram, K. and Whitt, W., "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data", *IEEE J. on Selected Areas in Communications* **4:6** (1986), 833–846.

[89] Steffenson, J.F., *Interpolation*, Chelsea Publishing Co., New York, 1950.

[90] Takács, L., *Introduction to the Theory of Queues*, Oxford University Press, New York, 1962.

[91] Takine, T., Sengupta, B. and Hasegawa, T., "An Analysis of a Discrete-Time Queue for Broadband ISDN with Priorities Among Traffic Classes". *IEEE Transactions on Communications* **42:2/3/4** (1994), 1837–1845.

[92] Taqqu, M.S. and Levy, J.B., "Using Renewal Processes to Generate Long-Range Dependence and High Variability", *Dependence in Probability and Statistics*, Eberlein, E and Taqqu, M.S. (Eds.), Progress in Prob. and Stat., Vol. 11, Birkhauser, Boston, 1986, 73–89.

[93] Taqqu, M.S., "Self-Similar Processes", *Encyclopedia of Statistical Sciences* **8**, Kotz, S. and Johnson, N. (Eds.), Wiley, new York, 1987.

[94] Wei, W.W.S., *Time Series Analysis: Univariate and Multivariate Methods*, Addison-Wesley, 1990.

[95] Whitt, W., "Tail Probabilities With Statistical Multiplexing and Effective Bandwidths in Multi-Class Queues", *Telecommunication Systems* **2** (1993), 71–107.

[96] Wilkinson, R., "Theories of Toll Traffic Engineering in the U.S.A.," *Bell System Technical Journal* **35:2** (1956), 421–514.

[97] Willinger, W., Taqqu, M.S., Leland, W.E. and Wilson, D.V., "Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements", *Statistical Science* **10** (1995), 67–85.

[98] Ye, J. and Li, S.Q., "Folding Algorithm: A Computational Method for Finite QBD Processes with Level-Dependent Transitions," *IEEE Trans. on Communications* **42:2** (1994), 625–639.