Statistical Multiplexing

Marwan M. Krunz
Department of Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85721
krunz@ece.arizona.edu

Conventional telecommunications networks, such as the public telephone network, were designed based on the synchronous transfer mode (STM) paradigm, which uses time-division multiplexing (TDM) for bandwidth allocation. In TDM, the link capacity is shared among contending connections using TDM *channels*. A channel is uniquely identified by the position of a time slot within a recurring synchronous structure, known as a frame. For two end systems to communicate, a logical connection must be established between them, whereby one or more STM channels are reserved for that connection. When a channel is assigned to a connection, its bandwidth cannot be shared with other connections (see Figure 1). The channel bandwidth is wasted when an established connection temporarily generates no traffic, as in the case of a listener in a phone conversation.

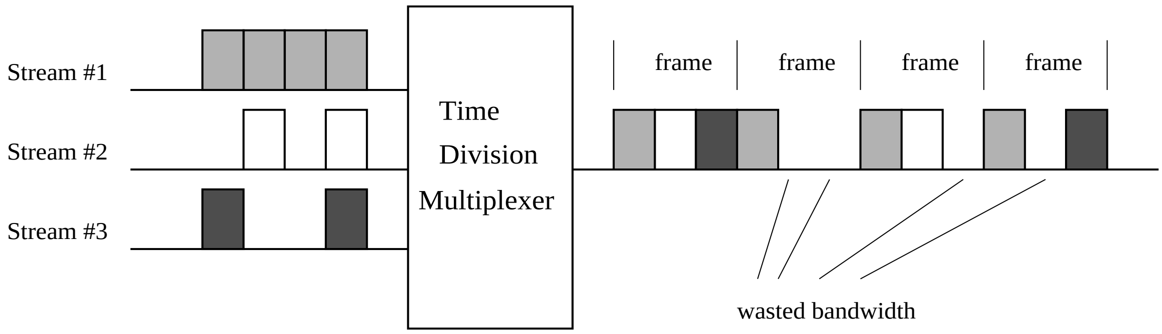

Figure 1: Time-division multiplexing (unused bandwidth is wasted).

In contrast to conventional networks, the architecture of Broadband-Integrated Services Digital Network (B-ISDN) is based on a new paradigm, known as the *asynchronous transfer mode* (ATM). The adoption of ATM as the transfer technology for B-ISDN came in response to several considerations. B-ISDN will offer the means of communications to a wide range of applications, including conventional voice and data applications as well as new multimedia applications (e.g., video-telephony, High-Definition TV, and multimedia conferencing). The integration of such diverse applications over a common communication platform requires a simple, unified transport technology, such as ATM, that is independent of the characteristics of the transported media. ATM is an attractive switching technology which is characterized by high-speed fiber transmission facilities and simple hardwired network protocols designed to match the huge transmission speeds of communication links. Transported data in ATM are encapsulated into fixed-length packets known as *cells*. The size of an ATM cell is 53 bytes; 5 bytes of which are used as a header. As a backbone switching network, ATM is designed to minimize the overhead incurred in processing network protocols. Cell switching in an ATM network is performed in hardware, unlike traditional packet-switched networks in which packets are routed by software processes.

1 Statistical Multiplexing

One important difference between ATM and STM is that instead of TDM, ATM uses *statistical* (or *asynchronous*) *multiplexing* (SM) as a means for resource sharing. In SM cells from various traffic streams share the link capacity on a need basis. Bandwidth is dynamically allocated so that if a stream is temporarily idle, its bandwidth is given to active streams. SM results in a significant improvement in bandwidth utilization, particularly when traffic streams are characterized by alternating active and idle periods with the active periods being, on average, shorter than the idle periods [86]. When statistically multiplexed, ATM connections are no longer identified by the location of a time slot in a synchronous structure. Instead, the header of each ATM cell contains connection identifiers that unambiguously identify the connection to which the cell belongs. This is illustrated in Figure 2 where three streams are statistically multiplexed onto an output link.

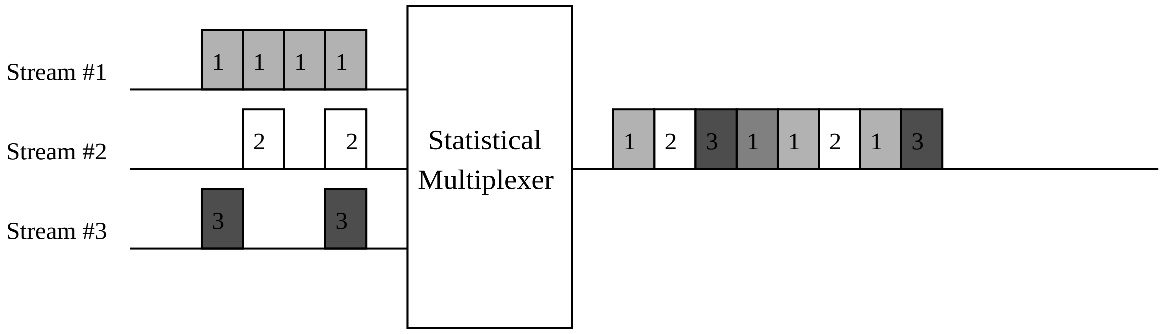

Figure 2: Statistical Multiplexing (bandwidth is allocated on demand).

A two-level hierarchy of connection identifiers is used in ATM networks: *virtual channel identifiers* (VCIs) and *virtual path identifiers* (VPIs). Each cell contains both identifiers in its header. A VCI specifies a Virtual Channel Connection (VCC). A VPI specifies a Virtual Path Connection (VPC), which is a bundle of VCCs. VCIs and VPIs have local scope that is limited to a given switch. During connection establishment, each switch along the path of a connection assigns VPI and/or VCI values to each connection, independently of the values assigned by other switches. When a cell arrives at a switch, the VPI and/or VCI in its header are checked. If the cell belongs to an admitted connection, the switch replaces the VPI and/or VCI in the header of the cell with the ones assigned to the connection. The cell is then sent to one of the output ports. Switching can be performed at the VP level with only VPIs being modified by a switch, or at the VC level with both VPI and VCI being modified. Typically, VCCs are statistically multiplexing at the output port of a switch.

2 Quality of Service in ATM Networks

Many multimedia applications require the underlying networking infrastructure to provide a priori guarantees on the transport of their packets. In the context of ATM, applications requirements are known as the *quality of service* (QoS), which is measured by throughput, cell loss, and cell delay metrics (including cell delay variation). Providing guarantees on the requested QoS is particularly important for interactive and real-time applications, where the timely delivery of packets is crucial to the coherent reception of the audio or video signal at the destination. In the connection establishment phase, applications provide their QoS requirements to the network. A connection is admitted only if the network can guarantee the requested QoS on an end-to-end basis without adversely affecting the QoS of already admitted connections.

2.1 Types of Network Guarantees

In principle, QoS guarantees can be offered on a deterministic or a statistical basis [53, 3, 91, 92]. Deterministic guarantees are hard bounds on the transport performance (for example, bounded packet transfer delay). Such bounds are relatively easy to support on an end-to-end basis, particularly in the case of the cell delay measure. However, deterministic guarantees are often provided at the expense of a conservative use of network resources, since the bounds are obtained under worst-case traffic assumptions.

While requiring some form of transport guarantees, some applications can tolerate infrequent violation of these guarantees. This is particularly true for "play-back" applications in which the traffic consists largely of audio-visual data units that are played back at the receiving end. Human insensitivity to small variations in picture and sound quality allows for a controlled loss of a small fraction of cells without any perceived impact on quality. For such applications, the network can provide statistical QoS. An example of such a guarantee is when the end-to-end cell transfer delay is ensured to be less than D_{max} with probability $1 - \alpha$, where $0 < \alpha \ll 1$ [53]. At steady-state, this means that $(1 - \alpha)\%$ of cells should encounter a delay of no more than D_{max}; the remaining cells which arrive late can be discarded by the receiver without any perceived degradation in the signal quality. Applications can compensate for some of the losses using error concealment mechanisms [65, 58]. The main advantage of supporting statistical guarantees is that they can be used in conjunction with statistical multiplexing, which significantly improves the utilization of network resources.

2.2 Approaches to Providing QoS Guarantees

Four approaches have been identified for providing QoS guarantees [53]. They represent different tradeoffs between simplicity and efficiency; a simple approach has practical appeal, but it often involves conservative allocation of resources.

2.2.1 Controlled Deterministic Approach

This approach provides deterministic QoS guarantees by shaping the traffic to conform to a predefined traffic envelope (a traffic envelope is a time-invariant deterministic bound on the bit rate). Some of the common envelopes include the (σ, ρ) model [17, 18], which has been extensively studied and has been used as the basis for the popular leaky-bucket traffic policing mechanism. The deterministic approach, while easy to implement, has two main disadvantages. First, traffic envelopes are inherently conservative, resulting in pessimistic performance predictions and poor bandwidth utilization. Second, because of the impact of statistical multiplexing on the characteristics of incoming streams, enforcing a particular envelope requires shaping the traffic after exiting each multiplexing node, which increases the hardware requirements of a switch.

2.2.2 Approximate Stochastic Approach

This approach is appropriate for applications that contend with statistical guarantees. Here, traffic streams are characterized by stochastic models that capture, to different degrees, the inherent randomness and fluctuations in the traffic. Statistical guarantees are provided by analyzing the performance of the multiplexer as a queueing system. The multiplexer is modeled as a queueing system of one or more finite-capacity queues that are served by a common server. A queue in this context is used as a surrogate to a memory buffer that accommodates arriving cells and queues them for switching onto the output link. The service rate is given by the transmission rate of the link. An ATM cell that arrives at the multiplexer is served immediately if the server is idle, or is queued for service if another cell is being served. The problem of providing guarantees can be formulated as follows: given a number of traffic streams that are statistically multiplexed, and given a stochastic model that characterizes the individual streams or their aggregate, find the bandwidth and buffer resources that must be allocated to these streams so that a given set (or sets) of statistical QoS are guaranteed. The statistical guarantees are obtained from the buffer overflow probability and the probability distribution function for the delay in the queue. The former gives an indication of the cell loss rate, while the latter can be used to approximately obtain various cell delay measures at the node. While the end-to-end delay consists of propagation, transmission, and queueing delays, only the queueing delay is variable and must thus be analyzed. One major problem with the approximate approach is that different traffic models give rise to different queueing behaviors. Some sophisticated models are sufficiently accurate, but their queueing analysis is not analytically tractable.

2.2.3 Bounding Stochastic Approach

Instead of employing detailed stochastic models, this approach contends with stochastic bounds on the number of arrivals in any interval of time of length T, possibly for several values for T [54]. This approach is the stochastic counterpart of the deterministic approach, where the traffic envelope here is specified in probabilistic terms. Aside from the bounds, no assumptions are made on the actual arrival pattern. The end-to-end guarantees are obtained by first obtaining stochastic bounds on the traffic at the edge of the network, which are then used to bound the departure traffic at that node. In turn, the bounded departure traffic of one node is used to bound the arrival traffic at the next node, and the procedure is repeated for all nodes along the path. A drawback of this approach is that the bounds become loose as more nodes are traversed.

2.2.4 Observation-Based Approach

In contrast to the previous three approaches, the observation-based approach [14, 44] provides no *a priori* quantitative commitments on performance. It rather uses online measurements to determine the current bandwidth demand and the admissibility of a new connection under given QoS requirements. The guarantees are thus "predictive" based on the network status at the time of connection establishment. For this reason, advocates of this approach prefer the term "predictive service" to indicate the qualitative nature of the guarantees. While the observation-based approach is probably the simplest among the four approaches, its qualitative nature precludes its use with non-adaptive applications that require a priori, quantitative commitments on performance.

3 Priority Mechanisms and Scheduling at a Multiplexer

Traffic streams transported over B-ISDN are expected to have a wide range of QoS requirements. Not only is this true for a heterogeneous mix of traffic, but also for certain individual traffic sources that generate cells with several loss and/or delay requirements, as in the case of MPEG-coded video, which uses layered coding to provide basic and enhanced layers of compressed video (information in the base layer is more crucial to the reconstruction of the video signal). Ideally, the network must guarantee the QoS of all connections while, simultaneously, taking advantage of statistical multiplexing. To do that, the network may choose to provide indistinguishable transport service based on the most stringent QoS requirements. Such a strategy is too restrictive and significantly underutilizes network resources, particularly when the traffic streams with the most demanding requirements constitute a small fraction of the total traffic. Alternatively, the network can be designed to offer multiple bearer capabilities by assigning levels of 'delivery' priority to incoming cells and offering differential service to these cells using priority queueing mechanisms. Such mechanisms can be implemented at various buffering stages in the network. The use of priority gives the network the flexibility to dynamically adjust to different traffic mixes, resulting in an increase in the total admissible load as compared to non-prioritization [48]. Priority mechanisms are also useful in other areas of traffic control such as traffic policing.

3.1 Types of Priority Mechanisms

In general, the design of a priority mechanism involves two aspects: a service (or scheduling) discipline, which determines the order in which cells in the buffer are served, and a buffer access discipline, which deals with admitting cells to the buffers [61]. Explicit or implicit priority rules may be applied to either or both disciplines. Accordingly, two types of priority queueing mechanisms can be identified, based on *where* the priority rule is enforced: delay and loss priority mechanisms.

3.1.1 Delay Priority Mechanisms

In a delay priority mechanism, the priority rule takes place at the output of the buffer. It is in essence a scheduling algorithm, with higher priority cells receiving preferential service over lower priority cells in the scheduling order. Delay priority mechanisms are quite useful for time-critical traffic, such as alarms and real-time control messages in manufacturing environments. Examples of delay priority mechanisms are Head-Of-the-Line (HOL), Earliest-Deadline-First (EDF), Queue-Length Threshold (QLT), Minimum-Laxity Threshold (MLT), and HOL with Priority Jumps (HOL-PJ) [11, 60, 10, 3]. A delay priority scheme can be static or dynamic. With the former type, the priority rule does not adapt to changes in the traffic mix or load conditions. In contrast, priority levels in a dynamic priority scheme are adjusted dynamically to cope with traffic conditions. Both QLT and MLT are of this type.

3.1.2 Loss Priority Mechanisms

The priority rule in this case is applied at the input to the buffer. Cells of higher classes have priority over cells of lower classes in terms of accessing the buffer. A loss priority mechanism is, therefore, a selective cell discarding scheme in which a cell of a given class is dropped (rather than delayed) to accommodate a higher priority cell. Loss priority mechanisms were first introduced to control congestion in ATM nodes [4]. They are needed to protect an ATM node from the stochastic fluctuations in the traffic which may temporarily deplete network resources and cause congestion to develop. These schemes are also used to guarantee different cell loss rates for various classes of traffic. The use of a loss priority scheme results in an increase in the total admissible load compared to no priority. Examples of loss priority mechanisms are Push-Out (PO) [29] and Partial Buffer Sharing (PBS) [26, 48, 49, 77]. In PO, cells enter one shared buffer up to the maximum buffer size. If a high priority cell arrives at a saturated buffer that contains low priority cells, a low priority cell is dropped and its place is given to the high priority cell. Despite its efficiency, PO requires a complicated buffer management to preserve the sequencing of cells. PBS achieves loss priority by means of threshold-based cell discarding. We now describe a generalized form of it, known as Nested-Threshold Cell Discarding (NTCD) [88, 77, 87, 49]. Under NTCD (see Figure 3) the buffer is partitioned by n thresholds, $T_1, \ldots, T_n$, that correspond to $n+1$ priority classes. Cells of priority class i enter the buffer up to threshold level T_i. Once the buffer level is above T_i, arriving cells of class i are dropped. Note that only new arrivals are dropped; class-i cells that are already in the buffer are never dropped and are eventually served. NTCD results in a slightly less total admissible load compared to PO [48], but it is less complex to implement in hardware.

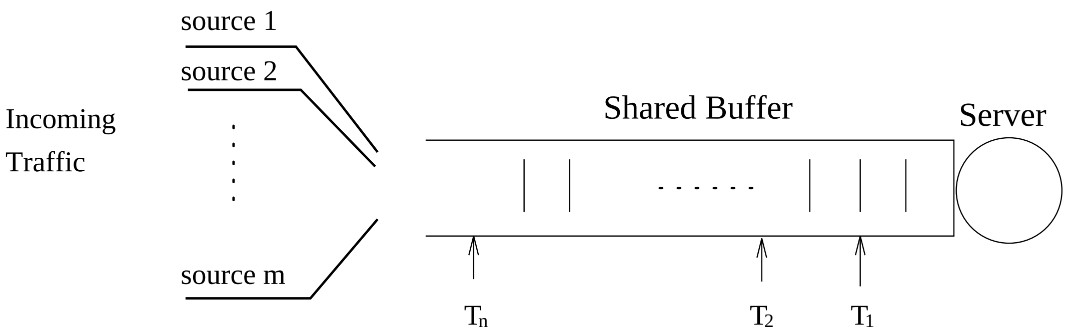

Figure 3: Example of a loss priority mechanism at a multiplexer: NTCD with n thresholds.

4 Burstiness and Traffic Correlations

Burstiness is an important characteristic of ATM traffic that has profound impact on the multiplexing performance. In simple terms, burstiness indicates the presence of non-negligible positive correlations between cell interarrival times. It arises naturally as a consequence of segmenting variable-length packets at the sender into fixed-length cells that are injected into the network. After segmentation, a traffic stream looks like a sequence of alternating active (ON) and idle (OFF) periods where each ON period consists of a "train" of cells following each other. Therefore, even if packets interarrival times are assumed to be uncorrelated, cells interarrival times are strongly correlated. The discovery of the bursty nature of ATM traffic was a turning point in the study of network traffic. Traditional renewal models, including the Poisson model, are no longer adequate as they tend to severely underestimate the queueing behavior at the multiplexer [81]. And even if the traffic of a single source is approximated by a renewal process, the traffic resulting from the superposition of finite number of sources is a complex non-renewal process that is modulated (i.e., controlled) by the number of active sources at each instant [34]. Correlations between arrivals have been found to cause considerable degradation in network performance (as measured by cell loss rate and delay jitter), which can not be predicted by a simple renewal model. The term "burstiness" has been traditionally used to indicate a correlated process in which the variance of the interarrival times is greater than the variance of the interarrival times of a Poisson process [33]. However, burstiness is better quantified by other measures, including the ones described below.

4.1 Indices of Dispersion

The index of dispersion for intervals (IDI) and the index of dispersion for counts (IDC) are two popular measures of burstiness [81, 33, 84, 1]. Let $\{X_k,\ k \geq 1\}$ be a sequence of interarrival times of a stationary arrival process. Let $S_k = X_1 + X_2 + \cdots + X_k$, for all k. The IDI is defined by the sequence $\{c_k^2,\ k \geq 1\}$ [16], where

$$c_k^2 = \frac{k\mathrm{Var}\,(S_k)}{[E\,(S_k)]^2} = \frac{\mathrm{Var}\,(S_k)}{k\,[E\,(X_1)]^2} = c_1^2 + \frac{\sum_{i,j=1,\, i\neq j}^{k} \mathrm{cov}\,(X_i,\ X_j)}{k\,[E\,(X_1)]^2}, \quad k \geq 1 \tag{1}$$

and $\mathrm{cov}\,(X_i,\ X_j)$ is the covariance of X_i and X_j. For $k > 1$, c_k^2 measures the cumulative covariance (normalized by the square mean) among k consecutive interarrival times of a source. The significance of the IDI measure is related to the fact that the multiplexing performance is influenced by the *cumulative* effect of covariances, rather than the individual covariances [81].

Let $N(t)$ be the counting process associated with $\{X_k\}$:

$$N(t) \triangleq \sum_{k=1}^{\infty} \mathbf{1}[S_k \leq t] \tag{2}$$

where $\mathbf{1}[.]$ is the indicator function. The IDC is defined by the function

$$I(t) = \frac{\mathrm{Var}[N(t)]}{E[N(t)]},\ t > 0 \tag{3}$$

For a Poisson process, $I(t) = c_k^2 = 1$ for all t and k. For a renewal process, $c_k^2 = c_1^2$ for all k. Accordingly, one can test the appropriateness of the renewal assumption in a modeling problem by examining the IDI of the modeled data.

Typically, one is interested in evaluating the multiplexing performance for a finite number of sources. In this case, it is more appropriate to consider the IDI and IDC for the *aggregate* traffic resulting from the superposition of several sources. Let $c_k^2(n)$ and $I(t,n)$ be the IDI and IDC for the superposition of n processes, respectively. Then, for mutually independent and identical renewal processes we have [16]:

$$\lim_{t\to\infty} I(t,n) = \lim_{k\to\infty} c_k^2(n) = c_1^2(1) \tag{4}$$

for any fixed n. The right equality says that for a fixed number of streams, as more interarrival times are considered, the IDI tends to the coefficient of variation (ratio of variance to mean) of a single stream. This interesting result points clearly to the inadequacy of Poisson modeling, since in practice $c_1^2(1) \gg 1$. Note, however, if $n \to \infty$ and the mean arrival rate of each stream is λ/n (i.e., individual processes become "sparse" as more streams are added), then the superposition process tends to a Poisson process (the Palm Theorem [13]). In practice, n is finite and the mean arrival rate is almost constant, independent of n.

4.2 Peakedness

Peakedness is another statistical measure of burstiness that was first used by teletraffic engineers to estimate call blocking probability at trunk groups [19]. Consider a stationary point process with rate λ. Each point corresponds to the arrival of a customer (a cell in the context of ATM). Let $\{N(t), t > 0\}$ be the counting process associated with the arrival process. Arrivals are offered to a group of infinite servers with *i.i.d.* service times and common service distribution G, which is also independent of the arrival process. Each customer is handled by its own server. Let $B(t)$ be the number of busy servers at time t. The peakedness of the arrival process with respect to a service distribution G is defined as

$$P_G = \lim_{t\to\infty} \frac{\mathrm{Var}[B(t)]}{E[B(t)]} \tag{5}$$

The peakedness of an arrival process is sometimes defined with respect to a family of service distributions that have the same form but differ in the value of one parameter, typically the mean. In this case, peakedness is indicated as a function of that parameter. For example, $P_{exp}(\mu)$ indicates the peakedness of an arrival process with respect to an exponential service distribution with mean μ. The peakedness has an apparent similarity to the IDC (see Equation 3). In fact, the limit of the IDC function can be expressed in terms of the peakedness with respect to an exponential distribution [67]:

$$\lim_{t\to\infty} I(t) = 2P_{exp}(0^+) - 1 \tag{6}$$

It is possible to express peakedness in terms of the second-order statistics of the arrival process and the service distribution. Let $\rho(t)$ be the autocorrelation function for the complementary service distribution:

$$\rho(t) \triangleq \int_0^\infty G^c(x)G^c(t+x)dx \tag{7}$$

where $G^c(x) = 1 - G(x)$. Then peakedness can be written as [19]:

$$P_G = 1 + \frac{\mu}{\lambda}\int_{-\infty}^{\infty}[k(x) - \lambda\delta(x)]\rho(x)dx \tag{8}$$

where $\delta(x)$ is the Dirac delta function and $k(x)$ is the covariance density of the arrival process, which is defined by

$$\frac{\partial^2 \mathrm{cov}[N(u), N(v)]}{\partial u \partial v} = k(u-v) = k(v-u), \quad \text{where } u,\ v > 0 \tag{9}$$

The above definition of peakedness applies only to point processes. Mark et al. [67] extended the definition to include fluid processes and batch arrivals. The new measure, known as *modified peakedness*, relies on the concept of a *rate process* $\{R(t) : t \in \mathbb{R}\}$, where $R(t)dt$ represents the amount of work (i.e., volume of arrivals) in the interval $[t, t+dt)$. The counting process can now be expressed as $N(t) = \int_0^t R(\tau)d\tau$, for $t > 0$. The work is offered to a fictitious service system that is characterized by a *service process* $\{U(t),\ t > 0\}$, where $U(t)$ can be interpreted as the time the work $R(t)dt$ spends in the system (the equivalent of the service time). The continuous-time process $\{U(t),\ t > 0\}$ is stationary with marginal distribution G. For any positive t_1 and t_2 with $t_1 \neq t_2$, $U(t_1)$ and $U(t_2)$ are *i.i.d.* random variables. The busy-server process is $\{\widetilde{B}(t),\ t > 0\}$, where

$$\widetilde{B}(t) \triangleq \int_0^t \mathbf{1}_{[U(x)>t-x]}(x)R(x)dx \tag{10}$$

Modified peakedness is defined by:

$$\widetilde{P}_G \triangleq \lim_{t\to\infty} \frac{\mathrm{Var}[\widetilde{B}(t)]}{E[\widetilde{B}(t)]} \tag{11}$$

Specifying the arrival process in terms of a rate process makes it possible to define the modified peakedness measure for processes that do not have the property of point arrivals, such as fluid processes and processes with batch arrivals.

5 Performance Analysis of a Statistical Multiplexer

Providing QoS guarantees necessitates analyzing the multiplexing performance at various locations in the network and computing the resources (bandwidth and buffer) needed to attain a certain level of QoS. Ideally, the performance is evaluated using measurements taken from an operational ATM network. At this point, ATM is still evolving and it has not yet been deployed at a wide scale. In the absence of an Internet-like ATM network, studying the performance of ATM multiplexers is typically done by means of analysis or, when analysis is intractable, by simulations. In either case, the multiplexer is modeled as a queueing system. Its input traffic is characterized by some stochastic process or by a set of "real" traces representing actual traffic patterns captured from an experimental ATM testbed. When real traces are used, the queueing performance is studied by means of discrete-event simulations, and the approach is known as *trace-driven simulations*. When a stochastic model is assumed, the queueing performance is often obtained analytically. However, some stochastic models do not lend themselves to tractable queueing analysis. Such models can still be used to generate synthetic traces (realizations of the underlying model), and the multiplexing performance is then evaluated by means of trace-driven simulations. The simulation-based approach has the disadvantage that it does not provide online results during connection establishment. Nonetheless, it can be useful, for example, when resources are to be dimensioned offline to guarantee a predetermined level of QoS. Moreover, a simulation-based approach separates the issues of traffic modeling and queueing evaluation, allowing highly accurate models to be employed (even if these models cannot be studied in an analytical queueing framework). It should also be mentioned that the queueing results for some traffic models, while analytically tractable, are not given in closed form. Thus, numerical computations must be performed to determine the measures of interest, such as the loss rate. These computations can be quite expensive, precluding the approach from being used in online admission control.

6 Traffic Models in ATM Networks

We now discuss some of the traffic models that have been used in studying the performance of an ATM multiplexer. Some of these models are more appropriate than others for a given type of traffic. The vast majority of traffic models are stochastic in nature, so they can only be used to provide statistical guarantees. Deterministic traffic models, which are not discussed here, include the (σ, ρ) model [17, 18], the D-BIND model [46], and other envelope-based models.

6.1 Renewal Models

Historically, queueing systems have been analyzed under renewal traffic models in which the interarrival times are *i.i.d.* A well-known example of renewal models is the Poisson process, in which the interarrival times are exponentially distributed. Renewal models include the so-called phase-type renewal processes [25], in which the interarrival times are derived from a continuous-time Markov process with discrete state space $\{0, 1, \ldots, M\}$. State 0 is absorbing, while all other states are transient. Starting with some probability distribution, an interarrival time is taken as the time to reach absorption. Subsequent interarrival times are obtained similarly by restarting the chain with the same initial distribution.

Interest in renewal models stemmed from their simplicity and analytical tractability. However, given the burstiness and the inherent correlations in ATM traffic, renewal models significantly underestimate the queueing performance, which is greatly affected by traffic correlations.

6.2 Markov and Markov-Modulated Models

6.2.1 Preliminaries

To account for traffic correlations, Markovian stochastic processes, which exhibit correlated interarrival times, have been extensively studied. Let $\{S_t : t \in \mathbb{R}\}$ be a continuous-time stochastic process with a sample space Ω. Then, $\{S_t\}$ is a Markov process if

$$\Pr\{S_{t_1} \leq x_1 / S_{t_2} = x_2, S_{t_3} = x_3, \ldots S_{t_n} = x_n\} = \Pr\{S_{t_1} \leq x_1 / S_{t_2} = x_2\} \tag{12}$$

for any $t_1 > t_2 > \cdots > t_n$ and any $x_1, \ldots, x_n$ in Ω. If the state space is discrete, the process is called a Markov chain and is associated with a transition probability matrix $\boldsymbol{P}$ that describes the probabilities of going from one state to another. The time spent in a given state is exponentially distributed with a parameter that depends on the current state. Typically, the transition probabilities are stationary, and $\boldsymbol{P}$ is expressed as $\boldsymbol{P} = [p_{ij}]$, where $p_{ij} \stackrel{\triangle}{=} \Pr\{S_n = x_j / S_{n-1} = x_i\}$; x_i and x_j are two states of the chain (discrete states are often taken as integers). When representing a traffic stream, transitions between states could represent cell arrivals.

6.2.2 Markov-Renewal Models

A Markov-renewal model consists of two processes: a Markov chain $\{S_n : n = 0, 1, \ldots\}$ and an associated transition-times process $\{T_n : n = 0, 1, \ldots\}$ [25]. At time n, the pair (S_{n+1}, T_{n+1}) of the next state depends only on the current state S_n. A transition from one state to another could indicate a cell arrival. This model has the advantage of allowing arbitrary interarrival times to be used, whereas only exponentially distributed interarrival times are possible in the basic Markov model.

6.2.3 Markovian Arrival Process (MAP)

MAPs are a subclass of Markov-renewal processes, which recently attracted much attention because of their versatility and analytical tractability [64]. As in phase-type renewal processes, interarrival times in a MAP are obtained from the time to reach absorption in a k-state Markov chain with one absorbing state and $k-1$ transient states. However, in contrast to a phase-type renewal process, the initial distribution for restarting the chain in a MAP depends on the last transient state from which the most recent absorption took place. This way interarrivals are correlated in a Markovian fashion. One important property of MAPs is that they obey a "superposition rule": The superposition of two independent MAPs is a MAP with an extended sample space. This property is quite useful in evaluating the performance of a statistical multiplexer. For example, if one traffic stream is modeled as a MAP, then the multiplexing performance for n such streams is given by the queueing performance under a single MAP with a larger state space. Extensions to MAPs include the batch MAP (BMAP) [62, 63] and the discrete-time BMAP (D-BMAP) [7].

6.2.4 Markov-Modulated Models

A modulated process is a doubly stochastic process whose parameters are modulated (controlled) by another stochastic process. Modulated processes play an important role in traffic modeling. Their versatility enables them to capture traffic randomness at multiple time scales. The simplest type of modulated processes uses a Markov process for modulation. Here, the probability law of the modulated process depends on the state of a modulating Markov chain. Each state gives rise to a different probability law. Typically, one parameter (e.g., the mean) is modulated. Popular Markov-modulated processes include the Markov-modulated Poisson process and Markov-modulated fluid models.

6.2.5 Markov-Modulated Poisson Process (MMPP)

The MMPP is a Poisson process whose arrival rate is a random variable that is modulated by the state of a continuous-time Markov chain. It is a correlated process that enjoys a tractable queueing analysis. One of its interesting features is that, similar to a MAP, the MMPP obeys a superposition rule: The superposition of two MMPPs is another MMPP with expanded state space [22]. An MMPP can be characterized by a generator matrix for the Markov chain and an associated arrival rate matrix.

Various MMPP-based traffic models have been used in the literature. One model uses a 2-state MMPP to characterize a single stream that alternates between active (ON) periods and idle (OFF) periods. In this case, the arrival rate during OFF periods is zero, and the MMPP reduces to an interrupted Poisson process (IPP). More commonly, an n-state MMPP (with $n \geq 2$) is used to characterize the *aggregate* of voice sources, each exhibiting an ON/OFF behavior [34] (the ON periods in a voice source correspond to talkspurts, while the OFF periods correspond to silence).

Various results related to the queueing performance under a MMPP arrival process are summarized in [22] (see also [89, 72, 88, 5]).

6.3 Fluid Models

The models discussed so far are all based on point processes, where the arrival of a cell is represented by a point on the time axis. A different approach to traffic modeling is provided by the fluid approximation [2, 47, 71]. Here, a traffic source is viewed as a stream of fluid that is characterized by a flow rate. The notion of discrete arrivals is lost as packets are assumed to be infinitesimally small (see Figure 4). The fluid approach has been found particularly appropriate to model the traffic in ATM networks for a number of reasons [20]. First, this approach captures the bursty nature of ATM traffic. Second, the traffic granularity, caused by small-size cells that are transmitted at very high speeds, makes the impact of individual cells insignificant. This gives a justification for the separation of cell-level and burst-level time scales, which is the underlying theme in the fluid approximation. Third, unlike the point process approach, the computational complexity of fluid analysis is independent of the buffer size, making the fluid approach particularly useful for systems with large buffers.

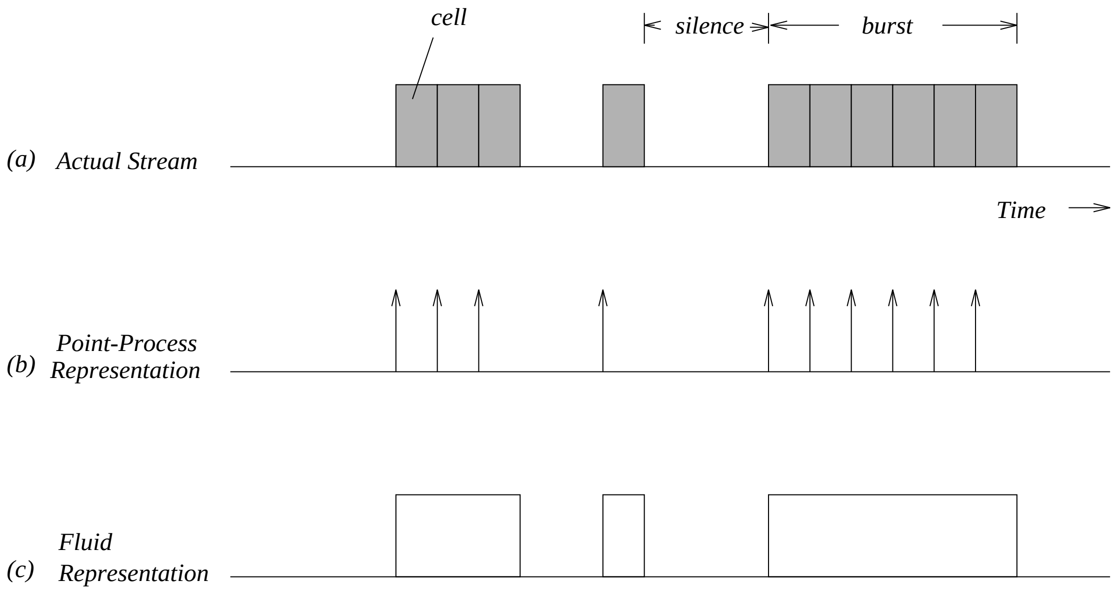

Figure 4: Point process and fluid representations of an ON/OFF traffic source.

Fluid models were originally developed for data and voice sources [2, 47]. When transmitted over a constant-rate line, bursty data and packetized voice streams (with silence detectors) exhibit the ON/OFF behavior. In the fluid model, the durations of the ON and OFF periods are random. During ON periods, the fluid arrives at a (constant) peak rate. ON periods (as well as OFF periods) are *i.i.d.*, often with exponentially distributed durations (the distribution for the ON periods is different from that of the OFF periods). Other scenarios have also been studied in the literature. The attractiveness of the exponential distribution is that it gives rise to a superposition rule that, in fact, applies to all Markov-modulated models (an important property of the exponential distribution is that the minimum of several independent and exponentially distributed rvs is another exponentially distributed rv). Consider the multiplexing of n homogeneous ON/OFF fluid sources; ON and OFF periods are exponentially distributed with means μ^{-1} and r^{-1}, respectively. Let λ be the arrival rate from one source during ON periods. The arrival process characterizing a single stream is a 2-state *Markov-modulated fluid flow* (MMFF) process, which is parameterized by a 2×2 infinitesimal generator matrix $\boldsymbol{Q}$ and an arrival-rate vector $\boldsymbol{\lambda} = (\lambda_0 \;\; \lambda_1) = (0 \;\; \lambda)$, where λ_i is the arrival rate in state i, $i \in \{0,1\}$. The superposition of the n streams is an $(n+1)$-state MMFF with generator matrix $\boldsymbol{Q}^{(n)}$ given by

$$\boldsymbol{Q}^{(n)} = \boldsymbol{Q} \oplus \boldsymbol{Q} \cdots \boldsymbol{Q} \; (n \text{ times}) \tag{13}$$

where $\oplus$ is the Kronecker sum. A state i in the Markov chain of the expanded MMFF means that i sources are simultaneously active (ON) and the remaining $n-i$ sources are idle. During state i, $i \in \{0,1,\ldots,n\}$, the total arrival rate is λi.

Fluid models enjoy tractable queueing analysis. In general, the queueing performance is obtained by formulating a set of first-order linear differential equations that describe the buffer occupancy at equilibrium in terms of the traffic parameters and the service rate. This set is then solved as a generalized eigenvalue/eigenvector problem (see [2, 82, 71] for details). Analytical results are also available for queues with priority scheduling: Elwalid and Mitra [20] analyzed a queue with multiple loss priorities and NTCD scheduling; Zhang [93] analyzed a two-buffer system in which one of the buffers has a complete preemptive priority (i.e., the multiplexer dedicates up to its full capacity to the high priority buffer, with the low priority buffer being served only when the high priority buffer is empty).

While the fluid approach is mathematically tractable, the queueing results are often obtained numerically (except for a few cases in which closed-form solutions are available). Unfortunately, the numerical procedure suffers from inherent numerical instability caused by the need to invert badly scaled matrices. Significant computations are needed to condition such matrices. The problem pertains to queues of finite capacity or queues that are partitioned by thresholds. Tucker [83] found a way to overcome the numerical problem, but his solution works only for a finite-capacity queue with no thresholds. Historically, the fluid approach used to suffer from a "state explosion" problem resulting from a large state space, but this problem was overcome in [82] by decomposition of the state space followed by a functional inversion.

In addition to ON/OFF sources, the fluid approximation was also used to model VBR sources generated by video-phones based on the conditional replenishment compression algorithm [66]. Let N_{vd} be the number of video sources arriving at a statistical multiplexer. In this model, the aggregate bit rate of the N_{vd} sources is quantized into a number of discrete levels, as shown in Figure 5. The quantized aggregate bit rate constitutes a Markov chain. Transitions between states

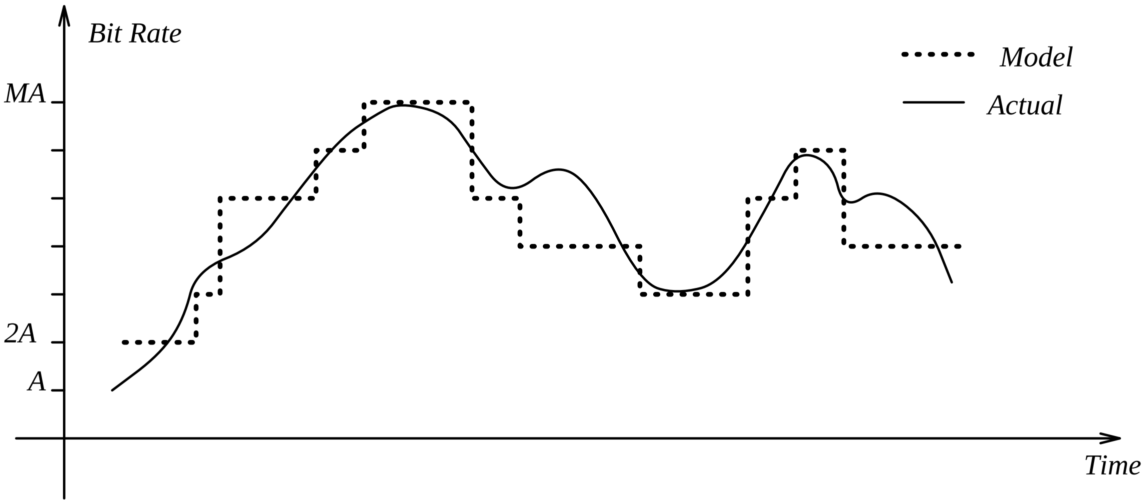

Figure 5: Quantization of the aggregate bit rate of multiplexed video-phone sources.

follow the birth-death transition diagram of Figure 6. At state i, the total arrival rate is $\lambda_{vd}(i) = Ai$, $i \in \{0, 1, \ldots, M\}$, where A is the quantization step (difference between two successive levels) and M is the number of quantization levels. Transitions occur only between adjacent states (i.e., arrival rate increases gradually). A transition from state i to state $i+1$ occurs at an average rate of $(M-1)r_{vd}$. Likewise, the average transition rate from state i to state $i-1$ is $i\mu_{vd}$. The process has the tendency to go to a lower level at high rates and to a higher level at low rates. The values for r_{vd} and μ_{vd} are found by matching the mean, standard deviation, and the autocorrelation function for the model and the empirical data.

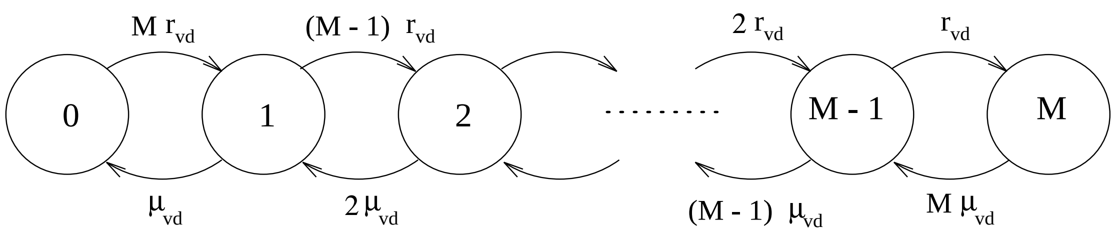

Figure 6: State transition diagram in the video-phone fluid model.

6.4 Regression Models

Regression models have been extensively used in fitting empirical time series arising in various domains, including finance, biological sciences, and engineering (cf. [8]). In teletraffic studies, regression models are found particularly suitable for characterizing compressed video streams. To maintain constant-quality video, a video encoder generates variable-size compressed frames at a constant frame rate (e.g., 30 frames/sec in the NTSC standard), so that the output stream has a variable bit rate. Characterizing the VBR stream is equivalent to modeling the frame-size sequence. Frame sizes are significantly affected by the scene dynamics. More dynamics means less temporal redundancy in the video, and thus larger encoded frames. The size of a frame is also influenced by the type of compression. Various compression techniques have been developed for video. These techniques vary in their compression efficiency, which usually comes at the expense of increased complexity in the encoder and decoder design, and therefore, higher encoding/decoding delay. Delay can be a deciding factor in the selection of a compression scheme. For example, interpolative motion compensation, which is part of the MPEG compression technique, is a very efficient compression scheme. However, its complexity prevents it from being used in real-time video conferencing.

Many regression models have been proposed for various types of video under different compression techniques. Earlier models were based on autoregressive (AR) and autoregressive moving-average (ARMA) processes [66, 38, 78, 90, 40]. An AR process of order p, AR(p), is a random process $\{X_n : n = 1, 2, \ldots\}$ that is given by:

$$X_n = a_0 + \sum_{r=1}^{p} a_r X_{n-r} + \epsilon_n, \quad n > 0 \tag{14}$$

where in the context of video, X_n models the size of the nth frame or, less commonly, a smaller unit than a frame, such as a *slice* (a horizontal strip in a frame). The sequence $\{\epsilon_n\}$ consists of *i.i.d.* random variables, known as the *residuals*, that give the AR model its stochastic nature. The residuals are often normally distributed, with mean zero, which implies that X_n is also normally distributed but with different mean and variance. In general, the autocorrelation function (ACF) of an AR(p) model is a mixture of damped exponentials and harmonics. It can be written as a difference equation [8]:

$$\rho_k = \sum_{r=1}^{p} a_r \rho_{k-r} \tag{15}$$

where ρ_k is the ACF at lag k (autocovariance at lag k divided by the variance). An AR(1) model was used in [66] to characterize the frame-size sequence of a video-phone under the conditional replenishment compression algorithm. For other types of video, higher-order AR models have been suggested, including AR(2) [38] and composite AR/Markovian models [78]. Autoregressive moving-average (ARMA) models have also been applied to the modeling of video streams [31]. For full-motion video, several regression models that explicitly incorporate scene dynamics have been investigated [36, 24, 37, 52]. In simple terms, a scene is a segment of a movie with no abrupt changes but possibly with some panning and zooming [55]. Frame sizes within a scene tend to be strongly correlated. To model scene dynamics, a discrete AR(1) (DAR(1)) process has been suggested [41, 38], in which frame sizes are generated according to a finite-state Markov chain. Once the chain enters a state, it stays there for a geometrically distributed random time, which corresponds to a scene length. The frame size stays constant during a scene but varies from one scene to another according to a negative binomial distribution [38] or a lognormal distribution [36, 51]. To provide better predictions of the queueing performance, Frater et al. [24] enhanced the DAR(1) model by using a different distribution for scene durations. As in the original DAR(1) model, bit-rate variations within scenes were not incorporated. An elaborate scene-based model for MPEG-coded movies was introduced in [52], which incorporates the three types of MPEG frames (I, P, and B). The model is composed of three sub-models, one for each frame type, which are intermixed in a deterministic, periodic manner. The sequence of I frames are modeled by the sum of two individually correlated processes: one process captures the variations between scenes, whose durations are assumed to be exponentially distributed, while the other is an AR(2) process that captures the intra-scene variations. Without the AR(2) component, the I-frame sub-model is simply the DAR(1) model with lognormal frame-size distribution. Two renewal processes were used to model the P and B sequences, with frame sizes having lognormal marginal distributions.

6.5 TES Models

Correlated random sequences can also be generated using the Transform-Expand-Sample (TES) technique [68, 42, 43], which is a form of nonlinear regression. The TES approach attempts to simultaneously capture the marginal distribution and the ACF of an empirical record. While the marginal distribution can be exactly matched, the ACF is approximated.

A TES model consists of two processes: a background process $\{U_n : n = 0, 1, \ldots\}$ and a foreground process $\{X_n : n = 0, 1, \ldots\}$. The background process defines a random walk on the unit circle (using modulo-1 arithmetic). It consists of a sequence of correlated and identically distributed random variables with a common distribution F_B. The foreground process is obtained by transforming the background process such that the resulting sequence has some target marginal distribution. In the general TES method, two transformations are needed to generate a foreground sequence with a target marginal distribution F:

$$X_n = F^{-1}(F_B((U_n)) \tag{16}$$

where F^{-1} (the inverse function) is known as the distortion function. The target distribution F is often expressed in the form of a histogram. Typically, F_B is a uniform distribution, and only one transformation is needed to generate the foreground sequence: $X_n = F^{-1}(U_n)$.

TES processes can be classified into TES$^+$ and TES$^-$, which differ in the sign of the ACF at lag 1. Let $\langle x \rangle$ indicate the modulo-1 of x (i.e., the fractional part). The background sequence in TES$^+$ is defined by

$$U_n^+ = \begin{cases} U_0, & \text{if } n = 0 \\ \langle U_{n-1}^+ + V_n \rangle, & \text{if } n > 0 \end{cases} \tag{17}$$

where U_0 is uniformly distributed on $[0, 1)$ and $\{V_n\}$ is an "innovation sequence" of *i.i.d.* random variables with marginal distribution F_V. The innovation sequence is independent of U_0. In TES$^-$ processes the background sequence is given by

$$U_n^- = \begin{cases} U_n^+, & \text{if } n \text{ is even} \\ 1 - U_n^+, & \text{if } n \text{ is odd} \end{cases} \tag{18}$$

Both types of background sequences can be shown to be Markovian with a uniform marginal distribution, irrespective of F_V. However, the transition probabilities for $\{U_n^-\}$ are time-dependent (i.e., $\{U_n^-\}$ is a non-homogeneous Markov process).

The modulo-1 operation results in sample paths that look more discontinuous at the origin than at other points. To make the sample path of a TES look more "homogeneous", a smoothing operation is applied prior to the transformation F^{-1}. This smoothing operation is called a stitching transformation and is given by the function:

$$S_\xi(x) = \begin{cases} x/\xi, & \text{if } 0 \le x < \xi \\ (1-x)/(1-\xi), & \text{if } \xi \le x < 1 \end{cases} \tag{19}$$

where ξ is called a stitching factor. The stitching transformation preserves the uniformity of the marginal distribution of the background sequence. The foreground sequence is now obtained using $X_n = F^{-1}(S_\xi(U_n)$, where U_n is either U_n^+ or U_n^-. Note that the transformation F_B is not needed since the background sequence consists of uniform variates.

The TES approach was used to model various types of ATM traffic, including H.261-encoded video [70], JPEG-encoded motion picture [56], and MPEG-encoded frame-level video [79]. In addition, a Markov-modulated TES process that accounts for scene dynamics was used to model JPEG-encoded motion picture [69]. The general approach to a TES model proceeds by searching for appropriate (ξ, F_V) that results in a sequence $\{S_\xi(U_n) : n = 0, 1 \ldots\}$ with an ACF that fits the ACF of the empirical data. Once this pair is determined, the transformation F^{-1} is applied, resulting in a correlated sequence with a desired distribution F. Because $\{S_\xi(U_n)\}$ has a uniform marginal distribution, the transformation F^{-1} preserves the correlation structure of $\{S_\xi(U_n)\}$. This way, the two aspects of fitting the ACF and the marginal distribution are decoupled.

One drawback of the TES approach is that the ACF of a TES process for lags beyond one cannot be given analytically. Thus, the pair (ξ, F_V) is obtained by systematically searching in the parameter space of ξ and F_V. There is no guarantee that the search will result in a good match to a given target ACF. Nonetheless, the TES approach is still one of the best available methods for generating correlated identically distributed random variates.

6.6 Long-Range Dependent Models

A common aspect of all the models presented so far is that the interarrival times are either uncorrelated or are correlated with an exponentially decaying ACF (e.g., Markovian models). Such models give rise to a summable ACF, i.e., $\sum_{k=0}^{\infty} \rho_k < \infty$. Recently, a number of studies supported by extensive measurements indicated the presence of persistent correlations in various types of network traffic, including LAN [23, 57] WAN [76], and VBR video traffic [6, 27]. This phenomenon, which is known as *long-range dependence* (LRD), has long been known in other domains of science, such as hydrolics and economics (cf. [6] and the references therein). It has been argued that the correlations persistence in network traffic cannot be adequately captured by Markov-like models. Instead, new models that exhibit the LRD behavior should be used to characterize network traffic and capture its correlations at multiple time scales.

The indication of the LRD phenomenon in network traffic spurred an ongoing debate on whether LRD models should be used in network dimensioning and resource allocation. The ramifications of LRD modeling are quite significant. For example, in contrast to Markovian models in which the burstiness is significantly tempered by statistical multiplexing, the multiplexing of LRD traffic streams can even increase traffic burstiness. While the persistence of traffic correlations is widely acknowledged, some researchers believe that capturing these correlations at all time scales is not needed to engineer the network. In particular, it is argued that since network buffers are finite in size, correlations beyond a certain critical lag have no impact on the queueing performance at a multiplexer [80, 35, 30].

6.6.1 LRD and Self-Similarity

Consider a second-order stationary process $\{X_n : n = 1, 2, \ldots\}$ with mean $\overline{X}$ and variance v. Let $C_k \triangleq \text{cov}(X_n, X_{n+k}) = E\left[(X_n - \overline{X})(X_{n+k} - \overline{X})\right]$. The ACF is given by $\rho_k = C_k/v$, for

$k = 0, 1, \ldots$. An equivalent representation to the ACF is given by its power spectral density:

$$g(\omega) \triangleq (1/2\pi) \sum_{k=-\infty}^{\infty} \rho_k e^{-ik\omega} \tag{20}$$

For $m = 1, 2, \ldots$, let $\{X_n^{(m)}\}$ be the time series obtained by averaging the original series $\{X_n\}$ over non-overlapping blocks of length m, i.e.,

$$X_n^{(m)} = \frac{1}{m}(X_{nm-m+1} + \cdots + X_{nm}), \quad \text{for } n = 1, 2, \ldots \tag{21}$$

The variance of the new time series is given by:

$$v_m \triangleq \text{var}(X_n^{(m)}) = \frac{v}{m} + \frac{2}{m^2} \sum_{p=1}^{m-1} \sum_{q=1}^{p} C_q \tag{22}$$

The process $\{X_n\}$ is said to be LRD if it satisfies any of the following (virtually equivalent) conditions [15]:

1. $\sum_{k=0}^{\infty} \rho_k = \infty$.

2. $g(w) \to \infty$ as $w \to 0$.

3. $mv_m \to \infty$ as $m \to \infty$.

If, on the other hand, $\sum_{k=0}^{\infty} \rho_k$ is finite, $g(0)$ is finite, or $lim_{j\to\infty} mv_m =$ constant, then $\{X_n\}$ is said to exhibit *short-range dependence* (SRD). Accordingly, all Markov-like models exhibit SRD. To exhibit LRD, the ACF of a model must drop off slowly, so that the autocorrelations have an infinite sum. Note that LRD is determined by the asymptotic behavior of the ACF (the sum of correlations up to a finite lag, no matter how large, does not determine whether or not the model is LRD). One particular form of LRD which received much attention is when:

$$\rho_k \sim ck^{-\beta} \text{ as } k \to \infty, \tag{23}$$

where $0 < \beta < 1$ and c is a constant. The slow decline of the power function results in a non-summable ACF.

Related to the LRD phenomenon is another interesting concept known as *self-similarity*, which in general terms states that the statistical properties of a stochastic process are invariant to the time scale. From (23), it can be shown that

$$\lim_{m\to\infty} \rho_k^{(m)} = (1/2)\delta^2(k^{2-\beta}) \approx \acute{c}k^{-\beta} \text{ (for large } k) \tag{24}$$

where $\rho_k^{(m)}$ is the autocorrelation in $\{X_n^{(m)}\}$ at lag k, $\delta^2(.)$ is the central second difference operator, and $\acute{c}$ is a constant. A process that satisfies (24) is said to exhibit *asymptotic* second-order self-similarity [15]. In contrast, a process exhibits *exact* second-order self-similarity if $\rho_k =$

$(1/2)\delta^2(k^{2-\beta})$ for all k, which also implies that $\rho_k^{(m)} = \rho_k$ for all non-negative integers k and m. Note that for SRD processes $\rho_k^{(m)} \to 0$ as $m \to \infty$, for all k, i.e., the process tends to white noise. In general, a process $\{X_t\}$ is said to be self-similar with parameter H, which is known as the *Hurst* parameter, if $\{X_{at} : t \geq 0\}$ and $\{a^H X_t : t > 0\}$ have identical finite-dimensional distributions for all $a > 0$. In other words, a self-similar process exhibits the same statistical properties (scaled by a^H) at all time scales. Interest is often limited to second-order self-similarity.

Of the several LRD models known in the literature, we will examine two important ones: the fractional autoregressive integrated moving-average (F-ARIMA) process and the fractional Guassian noise (FGN) process.

6.6.2 Fractional ARIMA Model

Long-range dependence is displayed by the F-ARIMA process, which is an extension of the conventional ARIMA processes [8]. An ARIMA(p, d, q) process $\{X_n\}$ is defined by:

$$\phi(B)\nabla^d X_n = \theta(B)\epsilon_n \tag{25}$$

where $\phi(B)$ and $\theta(B)$ are polynomials of orders p and q, respectively, in the delay operator B and ∇ is the differencing operator. In the F-ARIMA model, d is a fraction between 0 and 1/2. The F-ARIMA$(0, d, 0)$ model (i.e., $p = q = 0$) has been used to characterize VBR video streams [27, 39]. Letting $\phi(B) = \theta(B) = 1$, F-ARIMA$(0, d, 0)$ can be written as

$$\nabla^d X_n = \epsilon_n \tag{26}$$

The fractional differencing can be expanded as follows:

$$\nabla^d = (1-B)^d = \sum_{k=0}^{\infty} \begin{pmatrix} d \\ k \end{pmatrix} (-1)^k B^k \tag{27}$$

$$\begin{pmatrix} d \\ k \end{pmatrix} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)} \tag{28}$$

where $\Gamma(x) \triangleq \int_0^\infty t^{x-1}e^{-t}dt$ is the gamma function. When d is a positive integer, $\Gamma(d+1) = d!$. The ACF of the F-ARIMA$(0, d, 0)$ model behaves asymptotically as:

$$\rho_k = \frac{\Gamma(1-d)}{\Gamma(d)k^{2d-1}} \tag{29}$$

Thus, for $0 < d < 0.5$, the model exhibits LRD.

6.6.3 Fractional Guassian Noise Model

FGN is an exactly second-order self-similar process that is obtained from the stationary increments of a fractional Brownian motion (FBM). In itself, FBM is a self-similar Guassian process with Hurst parameter $H \in (0, 1)$. For the discrete-time case, the ACF of the (discrete) FGN is given by:

$$\rho_k = \frac{1}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right) \tag{30}$$

For $H \in (0.5, 1)$ $\rho_k \sim H(2H-1)k^{2H-2}$ as $k \to \infty$, and the FGN process is LRD.

6.7 $M/G/\infty$ Input Processes

$M/G/\infty$ processes constitute a versatile class of models that is capable of displaying many forms of correlations, including short-range and long-range dependence. They arise naturally as the limiting case for the superposition of many ON/OFF sources with ON periods having a "heavy-tailed" distribution such as the Pareto distribution [59]. Recently, $M/G/\infty$ processes have been proposed as a viable modeling approach for various types of network traffic [75, 74]. So far they have been applied in the modeling of VBR intra-coded video streams [50].

An $M/G/\infty$ process can be defined as follows: consider a discrete-time $M/G/\infty$ queue in which customers arrive in *i.i.d.* Poisson batches of mean λ. Let ξ_{n+1} be the size of the $(n+1)$th batch, i.e., the number of arrivals during time slot $[n, n+1)$. Upon arriving at the system, customers are presented to an infinite group of servers. Arrivals during time slot $[n, n+1)$ are serviced at the beginning of slot $[n+1, n+2)$. Let $\sigma_{n+1,1}, \ldots, \sigma_{n+1,\xi_{n+1}}$ be the integer-valued service times for customers $1, 2, \ldots, \xi_{n+1}$ of the $(n+1)$th batch, respectively. It is assumed that service times are *i.i.d.* with a common distribution G. We use σ to indicate a generic rv for the service time of a customer. Initially, there are b_{0^-} customers in the system with corresponding (residual) service times $\sigma_{0,1}, \ldots, \sigma_{0,b_{0^-}}$, which are mutually independent. Let b_n be the number of busy servers (i.e., remaining customers) at time n^+, $n = 0, 1, \ldots$, (after counting arrivals and departures at the start of slot $[n, n+1)$). The process $\{b_n : n = 0, 1, \ldots\}$ is known as the $M/G/\infty$ input process.

It has been shown that $\{b_n\}$ can display various forms of positive autocorrelations, the extent of which is controlled by the tail behavior of G [74]. In fact, it can even exhibit LRD when G is a Pareto distribution [15]. In general, $\{b_n\}$ is not stationary, but it admits a stationary and ergodic version $\{b_n^*\}$ [73], which is typically used as the basis for modeling. The ACF for the process $\{b_n^*\}$ is given by

$$\rho_k = e^{-u_k}, \quad \text{for } k = 0, 1, \ldots \tag{31}$$

where $u_k \stackrel{\triangle}{=} -\ln \mathrm{P}[\widetilde{\sigma} > k]$ and $\widetilde{\sigma}$ is the forward recurrence associated with the service time σ:

$$\mathrm{P}[\widetilde{\sigma} = i] = \frac{\mathrm{P}[\sigma \geq i]}{\mathrm{E}[\sigma]}, \quad i = 1, 2, \ldots \tag{32}$$

Equation (31) relates the monotonic behavior of the ACF to the service distribution G. By varying G, one can obtain various correlation structures. For example, a Weibull-like G was chosen in characterizing VBR video streams [50], so that the resulting ACF has the form $\rho_k = e^{-\beta\sqrt{k}}$, which provided a good fit to the empirical ACF. While this ACF is summable (i.e., the model is SRD), it does not exhibit a Markovian structure.

7 Effective Bandwidth

Statistical multiplexing improves network utilization by allowing bursty sources to share bandwidth on demand, so that the allocated bandwidth per source is less than the source peak rate. For the network to take advantage of statistical multiplexing, it should be able to approximately determine the minimum required bandwidth per source as a function of the QoS, buffer size at the multiplexer, and the traffic parameters. This bandwidth is commonly known as the *effective bandwidth* (or *equivalent capacity*). Once computed, the effective bandwidth can be used as the basis for call admission control (CAC). Intuitively, the effective bandwidth of a stream lies between the peak rate and the mean rate of that stream.

As a surrogate to a statistical multiplexer, consider an infinite-capacity queueing system with a single server and first-in-first-out (FIFO) service discipline. Input traffic consists of n independent, possibly heterogeneous, Markovian sources (e.g., fluid sources, MMPPs, MAPs, etc.). Let W be the waiting time of an arbitrary cell in the queue before it gets served. Under very general assumptions, the asymptotic behavior of the complementary distribution $G(x)$ for the waiting time is given by:

$$G(x) \stackrel{\triangle}{=} \mathrm{P}[W > x] \sim \alpha e^{\eta x}, \quad \text{as } x \to \infty \tag{33}$$

where η and α are called the *asymptotic decay rate* and *asymptotic constant*, respectively.

The literature provides several approximations for the effective bandwidth. The simplest of these is based on approximating (33) by

$$\mathrm{P}[W > x] \approx e^{\eta x} \tag{34}$$

(i.e., α is set to one) and using η as the basis for computing the effective bandwidth [21]. This one-parameter approximation, which we will refer to as the *asymptotic-rate-of-decay* (ARD) approximation, is quite appealing from a practical standpoint as it allows CAC to be designed solely based on basic characteristics of the input streams. Interestingly, according to the ARD approximation, the effective bandwidth of a source is independent of the characteristics of the other sources at the multiplexer. For general Markovian processes, the ARD approximation of the effective bandwidth of a source is determined by the maximal real eigenvalue of a matrix that is derived from the source parameters, network resources, and service requirements. Let p be the target overflow probability to be guaranteed by the network. For a buffer of size B, the QoS is satisfied if $G(B) \leq p$.

One special type of Markovian processes for which the ARD approximation was computed is given by Markov-modulated fluid flow (MMFF) processes. First, consider a multiplexer with only one input source, which is characterized by a MMFF process with generator matrix $\boldsymbol{M}$ (the infinitesimal generator matrix of the Markov chain) and arrival rates $\boldsymbol{\lambda} = (\lambda_1 \cdots \lambda_S)$, where λ_i is the fluid-flow rate during state i, $i \in \{1, \ldots, S\}$. Then, as $p \to 0$ and $B \to \infty$ with $\log p/B \to \xi \in [-\infty, 0]$, the ARD approximation of the effective bandwidth is given by the maximal real eigenvalue of the matrix $\boldsymbol{\Lambda} - (1/\xi)\boldsymbol{M}$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$. Now consider N multiplexed MMFF sources that are characterized by $(\boldsymbol{M}^{(j)}, \boldsymbol{\lambda}^{(j)})$, $j = 1, 2, \ldots, N$. The effective bandwidth for the superposition of these sources is given by the sum of their individual effective bandwidths which are given by the maximal real eigenvalue of the matrix $\boldsymbol{\Lambda}^{(j)} - (1/\xi)\boldsymbol{M}^{(j)}$, $j = 1, 2, \ldots, N$. The ARD approximation was also computed for MMPP sources, where the effective bandwidth is now given by the maximal real eigenvalue of the matrix $\frac{1}{e^{\xi}}\boldsymbol{\Lambda} - \frac{1}{1-e^{\xi}}\boldsymbol{M}$ (in this case, λ_i is defined as the average arrival rate of the Poisson process in state i).

According to the ARD approximation, the effective bandwidth for the aggregate traffic is simply the sum of the effective bandwidths for the individual sources, each computed independently of the other sources. This is a bit surprising since one would intuitively expect statistical multiplexing to reduce the effective bandwidth per source (after all, this is the purpose of statistical multiplexing). The unexpected result motivated further investigation of the effective bandwidth concept. It was found that the ARD approximation, though quite appealing for traffic management, is a conservative result, particularly in the moderate loss regimes (e.g., $p \geq 10^{-6}$). More specifically, the asymptotic constant, which is set to one in the ARD approximation, is found to decrease almost exponentially with n, the number of multiplexed streams [12]:

$$\mathrm{P}[W > x] \approx \alpha_n e^{-\eta x} \tag{35}$$

$$\alpha_n \sim \beta e^{\gamma n} \text{ as } n \to \infty \tag{36}$$

where $\beta > 0$. For sources more bursty than Poisson, $\gamma > 0$; otherwise, $\gamma \leq 0$.

A refined approximation of the effective bandwidth was developed in [12] based on the three-term approximation:

$$\mathrm{P}\left[W > x\right] \alpha_1 e^{-\eta_1 x} + \alpha_2 e^{-\eta_2 x} + \alpha_3 e^{-\eta_3 x} \tag{37}$$

where α_1 and η_1 are the asymptotic constant and asymptotic decay rate in (33). The other four parameters were chosen to match the probability of delay $\mathrm{P}\left[W > 0\right]$ and the first three moments of the delay. Other approaches to effective bandwidth estimation can be found in [32, 45, 9, 85, 28].

8 Open Issues and Challenges

While statistical multiplexing offers the means to achieve significant gain in network utilization, taking advantage of this gain necessitates a high degree of coordination among network elements. Protocols are needed to map the the end-to-end statistical QoS into nodal proportions that can by guaranteed by individual multiplexing nodes. Traffic models have been developed mainly for streams at entry nodes. To evaluate the multiplexing performance at intermediate nodes inside the network, the impact of statistical multiplexing on the input traffic at one node needs to be captured and incorporated in characterizing the departure traffic from that node. Video streams are often modeled under the assumption of uncontrolled bit rate. In practice, a connection is admitted based on a "traffic contract" that specifies the allowable values for the traffic descriptors of that connection (e.g., mean rate, peak rate, etc.). To prevent malicious or inadvertent violation of the contract, the network implements policing functions that enforce the contracted traffic descriptors. Estimating the appropriate traffic descriptors for VBR video is a challenging research issue that awaits further work.

References

[1] S. L. Albin. Approximating a point process by a renewal process, II: Superposition arrival processes to queues. *Operations Research*, 32:1133–1162, Sep.-Oct. 1984.

[2] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61(8):1871–1894, Oct. 1982.

[3] C. M. Aras, J. F. Kurose, D. S. Reeves, and H. Schulzrinne. Real-time communications in packet-switched networks. *Proceedings of the IEEE*, 82:122–139, Jan. 1994.

[4] J. J. Bae and T. Suda. Survey of traffic control schemes and protocols in ATM networks. *Proceedings of the IEEE*, 79(2):1871–1894, Feb. 1991.

[5] J. J. Bae and T. Suda. Analysis of a finite buffer queue with heterogeneous Markov modulated arrival processes: A study of the effects of traffic burstiness on individual packet loss. In *Proc. of IEEE INFOCOM '92*, pages 219–230, 1992.

[6] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-range dependence in variable bit-rate video traffic. *IEEE Transaction on Communications*, 43:1566–1579, 1995.

[7] C. Blondia and O. Casals. Statistical multiplexing of VBR sources: A matrix-analytic approach. *Performance Evaluation*, 16:5–20, 1992.

[8] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forcasting and Control*. Holden-Day, Inc., 1976. (Revised Edition).

[9] C.-S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. on Automat. Control*, 39:913–931, 1994.

[10] T. M. Chen, J. Walrand, and D. G. Messerschmitt. Dynamic priority protocols for packet voice. *IEEE Journal on Selected Areas in Communications*, 7(7), June 1989.

[11] R. Chipalkatti, J. F. Kurose, and D. Towsley. Scheduling policies for real-time and nonreal-time traffic in a statistical multiplexer. In *Proceedings of IEEE INFOCOM '89*, pages 774–783, 1989.

[12] G. L. Choudhury, D. M. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transaction on Communications*, 44(2):203–217, Feb. 1996.

[13] E. Cinlar. Superposition of point processes. In P. A. W. Lewis, editor, *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, pages 549–606. Wiley, 1972.

[14] D. D. Clark, S. Shenker, and L. Zhang. Supporting real-time applications in an integrated services packet network: Architecture and mechanism. In *Proceedings of the SIGCOMM'92 Conference*, pages 14–26, Aug. 1992.

[15] D. R. Cox. Long-range dependence: A review. In H. A. David and H. T. David, editors, *Statistics: An Appraisal*, pages 55–74. The Iowa State University Press, Ames, Iowa, 1984.

[16] D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events*. Methuen, London, England, 1966.

[17] R. L. Cruz. A calculus for network delay, part I: Network elements in isolation. *IEEE Transactions on Information Theory*, 37(1):114–131, Jan. 1991.

[18] R. L. Cruz. A calculus for network delay, part II: Network analysis. *IEEE Transactions on Information Theory*, 37(1):132–141, Jan. 1991.

[19] A. E. Eckberg. A generalization of peakedness to arbitrary arrival processes and service time distributions. *Bell Laboratories Technical Memorandum*, 1976.

[20] A. I. Elwalid and D. Mitra. Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic. In *Proceedings of IEEE INFOCOM '92*, pages 3c.4.1–3c.4.11, 1992.

[21] A. I. Elwalid and D. Mitra. Effective bandwidth for general Markovian traffic sources and admission control of high speed networks. *IEEE Journal on Selected Areas in Communications*, 1(3):329–343, June 1993.

[22] W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 1(18):149–171, 1993.

[23] H. J. Fowler and W. E. Leland. Local area network traffic characteristics, with implications for broadband network congestion management. *IEEE Journal on Selected Areas in Communications*, 9:1139–1149, 1991.

[24] M. R. Frater, J. F. Arnold, and P. Tan. A new statistical model for traffic generated by VBR coders for television on the Broadband ISDN. *IEEE Trans. on Circuits and Systems for Video Technology*, 4(6):521–526, Dec. 1994.

[25] V. S. Frost and B. Melamed. Traffic modeling for telecommunications networks. *IEEE Communications Magazine*, 32(3):70–81, Mar. 1994.

[26] J. Garcia and O. Casals. Priorities in ATM networks. In *Proc. of NATO Workshop on Architecture and Performance issues of high capacity local and metropolitan networks*, June 1990.

[27] M. W. Garrett and W. Willinger. Analysis, modeling, and generation of self-similar VBR video traffic. In *Proceedings of the SIGCOMM'94 Conference (Computer Communications Review)*, pages 269–280, Sept. 1994.

[28] R. J. Gibbens and P. J. Hunt. Effective bandwidths for the multi-type UAS channel. *Queueing Systems*, 9:17–28, 1991.

[29] A. Gravey and G. Hebuterne. Analysis of a priority queue with delay and/or loss sensitive customers. In *Proc. of ITC Broadband Seminar*, Oct. 1990.

[30] M. Grossglauser and J.-C. Bolot. On the relevance of long-range dependence in network traffic. In *Proceedings of the ACM SIGCOMM '96 Conference*, 1996.

[31] R. Grunenfelder, J. P. Cosmos, S. Manthorpe, and A. Odinma-Okafor. Characterization of video codecs as autoregressive moving average processes and related queueing system performance. *IEEE Journal on Selected Areas in Communications*, 9(3):284–293, Apr. 1991.

[32] R. Guerin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE Journal on Selected Areas in Communications*, 9(7):968–981, Sept. 1991.

[33] R. Gusella. Characterizing the variability of arrival processes with indexes of dispersion. *IEEE Journal on Selected Areas in Communications*, 9(2):203–211, Feb. 1991.

[34] H. Heffes and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):856–868, Sept. 1986.

[35] D. Heyman and T. Lakshman. What are the implications of long-range dependence for VBR video traffic engineering? Technical report, Bellcore, 1995.

[36] D. Heyman, E. Tabatabai, and T. Lakshman. Statistical analysis of MPEG2-coded VBR video traffic. In *Proc. of the Sixth International Workshop on Packet Video*, 1994.

[37] D. P. Heyman and T. V. Lakshman. Source models for broadcast video. *IEEE/ACM Transactions on Networking*, pages 40–48, Feb. 1996.

[38] D. P. Heyman, A. Tabatabai, and T. V. Lakshman. Statistical analysis and simulation study of video teleconferencing traffic in ATM networks. *IEEE Trans. on Circuits and Systems for Video Technology*, 2(1):49–59, Mar. 1992.

[39] C. Huang, M. Devetsikiotis, I. Lambadaris, and A. Kaye. Modeling and simulation of self-similar variable bit rate compressed video: A unified approach. In *Proc. of SIGCOMM '95*, 1995.

[40] B. Jabbari, F. Yegengolu, Y. Kuo, S. Zafar, and Y.-Q. Zhang. Statistical characterization and block-based modeling of motion-adaptive coded video. *IEEE Trans. on Circuits and Systems for Video Technology*, 3(3):199–207, June 1993.

[41] P. A. Jacobs and P. A. W. Lewis. Time series generated by mixtures. *Journal of Time Series Analysis*, 4(1):19–36, 1983.

[42] D. L. Jagerman and B. Melamed. The transition and autocorrelation structure of TES processes part I: General theory. *Stochastic Models*, 8(2):193–219, 1992.

[43] D. L. Jagerman and B. Melamed. The transition and autocorrelation structure of TES processes part II: Special cases. *Stochastic Models*, 8(3):499–527, 1992.

[44] S. Jamin, S. Shenker, L. Zhang, and D. Clark. An admission control algorithm for predictive real-time service (extended abstract). In *Third International Workshop on Network and Operating System Support for Digital Audio and Video*, pages 308–315, Nov. 1992.

[45] G. Kesidis, J. Warland, and C.-S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, 1(4):424–428, Aug. 1993.

[46] E. W. Knightly and H. Zhang. Traffic characterization and switch utilization using a deterministic bounding interval dependent traffic model. In *Proc. of IEEE INFOCOM '95*, pages 1137–1145, 1995.

[47] L. Kosten. Stochastic theory of data-handling systems with groups of multiple sources. *Performance of Computer-Communication Systems*, pages 321–331, 1984.

[48] H. Kroner, G. Hebuterne, P. Boyer, and A. Gravey. Priority management in ATM switching nodes. *IEEE Journal on Selected Areas in Communications*, 9(3):418–427, Apr. 1991.

[49] M. Krunz, H. Hughes, and P. Yegani. Design and analysis of a buffer management scheme for multimedia traffic with loss and delay priorities. In *Proc. of the IEEE GLOBECOM '94*, pages 1560–1564, San Francisco, CA, Nov. 1994.

[50] M. Krunz and A. Makowski. A source model for VBR video traffic based on $M/G/\infty$ processes. Technical Report CENG-TR-97-112, University of Arizona, Department of Electrical and Computer Engineering, 1997.

[51] M. Krunz, R. Sass, and H. Hughes. Statistical characteristics and multiplexing of MPEG streams. In *Proc. of the IEEE INFOCOM '95 Conference*, pages 455–462, Boston, Apr. 1995.

[52] M. Krunz and S. K. Tripathi. On the characterization of of VBR MPEG streams. In *Performance Evaluation Review (Proceedings of SIGMETRICS '97 Conference)*, June 1997.

[53] J. Kurose. Open issues and challenges in providing quality of service guarantees in high-speed networks. *Computer Communication Review, ACM/SIGCOMM*, 23(1):6–15, Jan. 1993.

[54] J. F. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *Proceedings of the ACM SIGMETRICS '92 Conference*, pages 128–139, June 1992.

[55] A. A. Lazar, G. Pacifici, and D. E. Pendarakis. Modeling video sources for real-time scheduling. Technical Report 324–93–03, Columbia University, Department of Electrical Engineering and Center for Telecommunications Research, Apr. 1993.

[56] A. A. Lazar, G. Pacifici, and D. E. Pendarakis. Modeling video sources for real-time scheduling. In *Proc. of IEEE GLOBECOM '93*, volume 2, pages 835–839, 1993.

[57] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, Feb. 1994.

[58] S. C. Liew and C. yin Tse. Video aggregation: Adapting video traffic for transport over broadband networks by integrating data compression and statistical multiplexing. *IEEE Journal on Selected Areas in Communications*, 14(6):1123–1137, Aug. 1996.

[59] N. Likhanov, B. Tsybakov, and N. Georganas. Analysis of an ATM buffer with self-similar fractal input traffic. In *Proceedings of IEEE INFOCOM '95*, pages 985–992, Boston, MA, Apr. 1995.

[60] Y. Lim and J. Kobza. Analysis of a delay-dependent priority discipline in a multiclass traffic packet switching node. In *Proc. of IEEE INFOCOM '88*, pages 889–898, 1988.

[61] A. Y.-M. Lin and J. A. Silvester. Priority queueing strategies and buffer allocation protocols for traffic control at an ATM integrated broadband switching system. *IEEE Journal on Selected Areas in Communications*, 9(9):1524–1536, Dec. 1991.

[62] D. M. Lucantoni. New results on a single server queue with a batch Markovian arrival process. *Stochastic Models*, 7:1–46, 1991.

[63] D. M. Lucantoni. The BMAP/G/1 queue: A tutorial. In L. Donatiello and R. Nelson, editors, *Models and Techniques for Performance Evaluation of Computer and Communications Systems*. Springer-Verlag, 1993.

[64] D. M. Lucantoni, K. S. Meier-Hellstern, and M. F. Neuts. A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22:676–705, 1990.

[65] W. Luo and M. El-Zarki. Analysis of error concealment techniques for MPEG-2 video transmission over ATM based networks. In *Proc. of SPIE/IEEE Visual Communications and Image Processing '95*, May 1995.

[66] B. Maglaris et al. Performance models of statistical multiplexing in packet video communications. *IEEE Journal on Selected Areas in Communications*, 36(7):834–844, July 1988.

[67] B. Mark, D. L. Jagerman, and G. Ramamurthy. Peakedness measures for traffic characterization in high-speed networks. In *Proceedings of IEEE INFOCOM '97*, 1997.

[68] B. Melamed. TES: A class of methods for generating autocorrelated uniform variates. *ORSA Journal on Computing*, 3(4):317–329, 1991.

[69] B. Melamed and D. Pendarakis. A TES-based model for compressed Star Wars video. In *Proceedings of IEEE GLOBCOM '94*, pages 120–126, 1994.

[70] B. Melamed, D. Raychaudhuri, B. Sengupta, and J. Zdepski. TES-based video source modeling for performance evaluation of integrated networks. *IEEE Transaction on Communications*, 42(10):2773–2777, Oct. 1994.

[71] D. Mitra. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability*, 20:646–676, 1988.

[72] R. Nagarajan, J. Kurose, and D. Towsley. Approximation techniques for computing packet loss in finite-buffered voice multiplexers. *IEEE Journal on Selected Areas in Communications*, 9(3):368–377, Apr. 1991.

[73] M. Parulekar. *Buffer Engineering for Self-Similar Traffic*. PhD thesis, University of Maryland, College Park, 1997.

[74] M. Parulekar and A. M. Makowski. Tail probabilities for M/G/∞ input processes (i): Preliminary asymptotics. Private Communications.

[75] M. Parulekar and A. M. Makowski. Tail probabilities for a multiplexer with self-similar traffic. In *Proceedings of IEEE INFOCOM '96*, pages 1452–1459, 1996.

[76] V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1993.

[77] D. W. Petr and V. S. Frost. Nested threshold cell discard for ATM overload control: optimization under cell loss constraints. In *Proc. of IEEE INFOCOM '91*, pages 12A.4.1–12A.4.1.10, Bal Harbour, 1991.

[78] G. Ramamurthy and B. Sengupta. Modeling and analysis of a variable bit rate video multiplexer. In *Proc. of IEEE INFOCOM '92*, volume 2, pages 817–827, 1992.

[79] D. Reininger, B. Melamed, and D. Raychaudhuri. Variable bit rate MPEG video: Characteristics, modeling and multiplexing. In *Proc. of the 14th Int. Teletraffic Congress*, pages 295–306, Antibes Juan-les-Pins, France, June 1994.

[80] B. Rye and A. Elwalid. The importance of long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities. In *Proceedings of the ACM SIGCOMM '96 Conference*, Aug. 1996.

[81] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications*, 4(6):833–846, Sept. 1986.

[82] T. E. Stern and A. I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Advances in Applied Probability*, 23:105–139, 1991.

[83] R. C. F. Tucker. Accurate method for analysis of a packet-speech multiplexer with limited delay. *IEEE Transaction on Communications*, 36(4):479–483, Apr. 1988.

[84] W. Whitt. Approximating a point process by a renewal process: Two basic methods. *Operations Research*, 30(1):125–147, 1982.

[85] W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidths for multiclass queues. *Telecommunications Systems*, 2:71–107, 1993.

[86] G. Woodruff and R. Kositpaiboon. Multimedia traffic management principles for guaranteed ATM network performance. *IEEE Journal on Selected Areas in Communications*, 8(3):437–446, Apr. 1990.

[87] V. Yau and K. Pawlikowski. ATM buffer overflow control: A Nested Threshold Cell Discarding with Suspended Execution. In *Proc. of Australian Broadband Switching and Services Symp.*, Melbourne, July 1992.

[88] P. Yegani. Performance models for ATM switching of mixed continuous bit rate and bursty traffic with threshold based discarding. In *Proc. of IEEE ICC '92*, pages 1621–1627, June 1992.

[89] F. Yegengolu and B. Jabbari. Performance evaluation of MMPP/D/1/K queues for aggregate ATM traffic models. In *Proceedings of IEEE INFOCOM '93*, pages 1314–1319, 1993.

[90] F. Yegengolu, B. Jabbari, and Y.-Q. Zhang. Motion-classified autoregressive modeling of variable bit rate video. *IEEE Trans. on Circuits and Systems for Video Technology*, 3(2):42–53, Feb. 1993.

[91] H. Zhang and S. Keshav. Comparison of rate-based service disciplines. In *Proceedings of the ACM SIGCOMM '91 Conference*, pages 113–121, Sept. 1991.

[92] H. Zhang and E. W. Knightly. Providing end-to-end statistical performance guarantees with bounding interval dependent stochastic models. In *Proc. of SIGMETRICS '94*, pages 211–220, 1994.

[93] J. Zhang. Performance study of Markov modulated fluid flow models with priority traffic. In *Proc. of IEEE INFOCOM '93*, pages 1a.2.1–1a.2.8, 1993.

Cross-References

Asynchronous Transfer Mode, Traffic Models