# PERFORMANCE OF TELECOMMUNICATION SYSTEMS: SELECTED TOPICS

K. Kontovasilis
*N.C.S.R. "Demokritos", Inst. Informatics and Telecommunications,*
*P.O. Box 60228, GR-15310 Aghia Paraskevi Attikis, Greece*

S. Wittevrongel, H. Bruneel
*Ghent University, Dept. Telecommunications and Information Processing*
*SMACS Research Group, St. Petersnieuwstraat 41, B-9000 Gent, Belgium*

B. Van Houdt, C. Blondia
*University of Antwerp, Dept. Mathematics and Computer Science*
*PATS Research Group, Universiteitsplein 1, B-2610 Antwerpen, Belgium*

## 1. Introduction

The central activity of performance evaluation is building formal descriptions of the system under study, an activity refered to as modelling. These models include workload models (e.g. packet traffic models), system resource models (e.g. switch models, link models) and resource control mechanism models (e.g. MAC protocol models). They are used to gain insight in the performance of the system under certain load conditions. To obtain the performance measures of interest, two technics exist: simulate the system (i.e. built a program that simulates the model behavior) or solve the model mathematically (i.e. compute the performance measures analytically). In this paper, we concentrate on the latter, making distiction between analytical methods that lead to closed formulas (as described in the part on generating functions) and algorithms that allow to compute the measures numerically (as in dealing with matrix analytic methods).

An area where performance modeling is an essential tool for system designers and developers today is the Internet. The Internet is evolving from a best-effort network towards a system that combines Quality of

L. Chapin (ed.), *Communication Systems*

Service (QoS) support with efficient resource usage. The extremely rapid pace of change that can be observed in the Internet research community (e.g. in the IETF), often does not allow rigorous performance evaluation of the different proposals. Therefore, the performance evaluation community (e.g. IFIP WG 6.3) should make an effort to provide the necessary methods, techniques and tools to allow a better insight into the system behavior.

The first part is devoted to the modeling of telecommunications sytems using generating functions, the second part to modeling using matrix analytical methods and finaly, the third part to the use of asymptotic approximations. The first part considers discrete-time queueing models as representatin of telecommunication systems. These models are particularly applicable when the time can be segmented in intervals of fixed lenght (called slots) and information packets are transmitted at slot boundaries. A typical example is an ATM transmission system. The method to compute the performance measures of interest is based on the use of generating functions. The aim is to obtain a closed form formula for the generating function of the system content (i.e. how many packets are present in the system), from which the most important performance measures can be derived. The main characteristic of this approach is that it is almost entirely analytical.

A second approach to compute performance measures is found in Section 3. Here two recent and promising developments within the framework of matrix analytical methods are discussed. Both models, have important applications in the performance analysis of telecommunication systems. The first model is concerned with a markovian arrival process with marked arrivals, of particular interest in systems where the packets are originating from different possiblly correlated traffic streams. A second model deals with Tree structured Markov chains. Their particular structure can be exploited to obtian efficient computational methods to obtain the measures of interest. Random access algorithms known as stack algorithms, or tree algorithms with free access, are examples of systems that can be modeled by means of tree stuctured Markov chain, leading to expressions for the maximum stable throughput and mean delay in such systems.

A third part is devoted to asymptotic approximations. These methods have become extremely relevant due to the high transmission rates and the stringent quality of service guarantees of modern systems, which make very rate events (e.g. buffer overflow) significant. Hence, performance measures are based on distribution tails, rather than on first moments. This paper studies both asymptotics for multiplexers with small buffers and for multiplexers with large buffers. In both cases, the

aim is to link the traffic load and resource capacity to the probability of loss due to buffer overflow. Also the case where the buffer space is neither negligible nor dominant is discussed.

# 2. Performance Modeling of Communication Systems Using Generating Functions

## 2.1. Discrete-time queueing models

In various subsystems of telecommunication networks, buffers are used for the temporary storage of digital information units which cannot be transmitted to their destination immediately. The performance of a communication network may be very closely related to the behavior of these buffers. For instance, information units may get lost whenever a buffer is fully occupied at the time of their arrival to this buffer, they may experience undesirable delays or delay variations in buffers, ... Queueing theory thus plays an important role in the performance modeling and evaluation of telecommunication systems and networks. In particular, queueing models in discrete time are very appropriate to describe traffic and congestion phenomena in digital communication systems, since these models reflect in a natural way the synchronous nature of modern transmission systems, whereby time is segmented into intervals ("slots") of fixed length and information packets are transmitted at slot boundaries only, i.e., at a discrete sequence of time points.

In a discrete-time queueing model, the arrival stream of digital information into a buffer (the input or arrival process) is commonly characterized by specifying the numbers of arriving packets during the consecutive slots. In basic models, these numbers of arrivals are assumed to be independent and identically distributed (i.i.d.) discrete random variables, and the corresponding arrival process is referred to as an independent or uncorrelated arrival process. More advanced models allow the numbers of arrivals during consecutive slots to be nonindependent, and are referred to as correlated arrival processes. The storage capacity of a buffer is usually modeled as unlimited. This is an acceptable assumption since in most communication systems the capacity is chosen in such a way that the loss probabilities are very small, and furthermore, this facilitates the use of analytical analysis techniques. The transmission of information units from the buffer (the output process) is characterized by the distribution of the transmission times of the information units, the number of output channels of the buffer, the availability of the output channels, and the order of transmission (the queueing discipline). In basic models, all information units are assumed to be of fixed length, which implies they have constant transmission times, the output chan-

nels are permanently available, and the queueing discipline is assumed to be first-come-first-served (FCFS). In some applications, however, it is necessary to consider non-deterministic transmission times, interruptions of the output channels, and non-FCFS queueing disciplines such as e.g. priority queueing.

In the next section, we present an overview of a number of fundamental techniques for the analysis - in the steady state - of a wide range of discrete-time queueing models. The main characteristics of the techniques are that they are almost entirely analytical (except for a few minor numerical calculations) and that an extensive use of probability generating functions is being made. Note that a steady state only exists if the mean number of packet arrivals per slot is strictly less than the mean number of packets that can be transmitted per slot.

## 2.2. Steady-state queueing analysis using generating functions

The behavior of a queueing system is commonly analyzed in terms of the probability distributions of the buffer contents, i.e., the total number of packets present in the buffer system, and the packet delay, i.e., the amount of slots a packet spends in the system.

**Buffer contents.** The first step in the analysis of the buffer contents is to establish a so-called "system equation" that describes the evolution in time of the buffer contents. If we define $s_k$ as the buffer contents at the beginning of slot $k$, it is easily seen that the following basic relationship holds :

$$s_{k+1} = s_k - t_k + e_k \ , \tag{1}$$

where $e_k$ represents the total number of packet arrivals during slot $k$ and $t_k$ denotes the number of packets that leave the buffer system at the end of slot $k$. Here the characteristics of $e_k$ depend on the specific nature of the arrival process. The random variable $t_k$ on the other hand depends on the characteristics of the output process, and cannot be larger than $s_k$ in view of the synchronous transmission mode, which implies that only those packets present in the buffer at the beginning of a slot are eligible for transmission during the slot.

In the simplest models, uncorrelated arrivals from slot to slot, constant transmission times of one slot each, and permanently available output channels are assumed. In this case, the system equation (1) reduces to

$$s_{k+1} = (s_k - c)^+ + e_k \ . \tag{2}$$

Here $(...)^+ = max(0, ...)$, $c$ denotes the number of output channels, and the random variables $s_k$ and $e_k$ on the right-hand side are statistically independent of each other, which implies that the set $\{s_k\}$ forms a Markov chain. Let $S_k(z) = E[z^{s_k}]$ denote the probability generating function (pgf) of $s_k$. By means of standard $z$-transform techniques [9], the system equation (2) can then be translated into the $z$-domain. This yields the following relationship between the pgf's $S_{k+1}(z)$ and $S_k(z)$ :

$$S_{k+1}(z) = E(z)\, z^{-c} \{ \sum_{j=0}^{c-1} (z^c - z^j) \,\mathrm{Prob}[s_k = j] + S_k(z) \} \ , \qquad (3)$$

where $E(z)$ denotes the pgf of the number of packet arrivals in a slot. In the steady state, both $S_{k+1}(z)$ and $S_k(z)$ will converge to a common limiting function $S(z)$, the pgf of the buffer contents $s$ as the beginning of an arbitrary slot in the steady state. Taking limits for $k \to \infty$ and solving the resulting equation for $S(z)$, we then obtain

$$S(z) = \frac{E(z) \sum_{j=0}^{c-1} (z^c - z^j) \,\mathrm{Prob}[s = j]}{z^c - E(z)} \ . \qquad (4)$$

The $c$ unknown constants $\mathrm{Prob}[s = j]$, $0 \le j \le c-1$, in (4) can be determined by invoking the analyticity of the pgf $S(z)$ inside the unit disk $\{z : |z| \le 1\}$ of the complex $z$-plane, which implies that any zero of the denominator of (4) in this area must necessarily also be a zero of the numerator, together with the normalization condition $S(1) = 1$ of the buffer-contents distribution. This results in a set of $c$ linear equations in the $c$ unknown probabilities and allows to obtain $S(z)$ explicitly.

During the last few years research has largely focused on the introduction of more complicated characterizations of the arrival process, in order to obtain more realistic, useful and tractable stochastic descriptions of the sometimes bursty and heterogeneous traffic streams occurring in modern integrated communication networks. When the arrival process is correlated, the random variables $s_k$ and $e_k$ on the right-hand side of the system equation (2) are no longer statistically independent, and the above analysis technique needs to be modified. Specifically, since the knowledge of the value of $s_k$ no longer suffices to determine the probability distribution of $s_{k+1}$, the set $\{s_k\}$ does no longer form a Markov chain, and a more-dimensional state description of the system has to be used, containing extra information about the state of the arrival process.

As an example, let us consider a discrete-time queueing model with one output channel, that is permanently available, and a simple corre-

lated arrival process. Packets are generated by $N$ independent and identical on/off-sources. Each source alternates between on-periods, during which it generates one packet per slot, and off-periods, during which no packets are generated. The successive on-periods and off-periods of a source are assumed to be independent and geometrically distributed with parameters $\alpha$ and $\beta$ respectively. Clearly, we then have the system equation (2), where $c = 1$, whereas $e_k$ can be derived from $e_{k-1}$ as follows [7] :

$$e_k = \sum_{i=1}^{e_{k-1}} c_i + \sum_{i=1}^{N-e_{k-1}} d_i \ . \tag{5}$$

Here the $c_i$'s and the $d_i$'s are two independent sets of i.i.d. Bernoulli random variables with pgf's

$$c(z) = 1 - \alpha + \alpha z \tag{6}$$

and

$$d(z) = \beta + (1 - \beta) z \ . \tag{7}$$

From (2) and (5)-(7), the pair $(e_{k-1}, s_k)$ is easily seen to constitute a (two-dimensional) Markovian state description of the system at the beginning of slot $k$. We then define $P_k(x, z)$ as the joint pgf of the state vector $(e_{k-1}, s_k)$, i.e.,

$$P_k(x, z) = E[x^{e_{k-1}} z^{s_k}] \ . \tag{8}$$

The next step is then similar to the uncorrelated-arrivals case, namely to derive a relationship between the pgf's $P_{k+1}(x, z)$ and $P_k(x, z)$ corresponding to consecutive slots, by means of the state equations :

$$\begin{aligned} P_{k+1}(x, z) &= E\left[(xz)^{e_k} z^{(s_k-1)^+}\right] \\ &= [d(xz)]^N E\left[\left(\frac{c(xz)}{d(xz)}\right)^{e_{k-1}} z^{(s_k-1)^+}\right] \\ &= \frac{[d(xz)]^N}{z} \left\{P_k\left(\frac{c(xz)}{d(xz)}, z\right) + (z-1)\,\mathrm{Prob}[s_k = 0]\right\} \ . \end{aligned} \tag{9}$$

Again taking limits for $k \to \infty$, we now obtain a "functional equation" for the limiting function $P(x, z)$, which typically contains the $P$-function on both sides, but with different arguments. Although the function $P(x, z)$ cannot be derived explicitly from the functional equation, several performance measures related to the buffer contents can be derived from it, as will be explained later.

A similar analysis is possible for a variety of correlated arrival processes, such as train arrivals [67], [65], Markov modulated arrivals [68],

[2], general on/off sources [64], correlated train arrivals [47], and so on. For some arrival processes, the resulting functional equation may contain a number of unknown boundary probabilities, which in general are difficult to obtain exactly. An approximation technique can then used, which is based on the observation that a buffer contents equal to $n$ at the beginning of a slot implies that no more than $n$ packets have entered the buffer during the previous slot (see e.g. [2], [64]).

Also in case more complicated models for the output process are used, similar problems occur and a more-dimensional state description needs to be used. For instance, when general transmission times are considered, additional information is needed in the state description about the amount of service already received by the packet(s) in transmission, if any [8]. In case interruptions of the output channels may occur, we need to keep track of the state of each of the output channels (available or blocked) and the remaining sojourn time in this state [15].

**Packet delay.** The delay of a packet is defined as the number of slots between the end of the slot of arrival of the packet, and the end of the slot when this packet leaves the buffer. In case of a FCFS queueing discipline, the analysis of the packet delay typically involves the derivation of a relationship between the delay of a tagged packet and the total number of packets present in the buffer just after the arrival slot of the tagged packet and to be transmitted before the tagged packet. However, for discrete-time queueing systems with one permanently available output channel, constant transmission times of one slot, a FCFS queueing discipline and an arbitrary (possibly correlated) arrival process, the following relationship exists between the pgf $S(z)$ of the buffer contents and the pgf $D(z)$ of the packet delay [59]:

$$D(z) = \frac{S(z) - S(0)}{1 - S(0)} \ . \tag{10}$$

The above relationship makes a full delay analysis superfluous, once the buffer contents has been analyzed. Similar relationships also exist in case of multiple servers [61] and non-deterministic service times [60].

## 2.3. Performance measures

The results of the analysis can be used to derive simple and accurate (exact or approximate) formulas for a wide variety of performance measures of practical importance, such as mean and variance of buffer occupancies and delays, packet loss probabilities, ... The mean system contents and the mean packet delay in the steady state can be found by evaluating the first derivative of $S(z)$ and $D(z)$ at $z = 1$. Higher-order

moments of the system contents and the packet delay can be derived analogously, by calculating higher-order derivatives of $S(z)$ and $D(z)$ at $z = 1$. The tail distribution of the buffer contents is, for reasons of computational complexity, often approximated by a geometric form based on the dominant pole $z_0$ of the pgf of the buffer contents. That is, for large values of $n$, the tail distribution of the buffer contents is approximated by [26]

$$\text{Prob}[s = n] \approx -\frac{\theta}{z_0} z_0^{-n} \ , \tag{11}$$

where $\theta$ is the residue of $S(z)$ for $z = z_0$. A quantity of considerable practical interest is the probability that the buffer contents (in the infinite buffer) exceeds a given threshold $S$. This probability can be used to derive an approximation for the packet loss ratio (i.e., the fraction of packets that arrive at the buffer but cannot be accepted) of a buffer with finite waiting space $S$ and the same arrival statistics [50].

As mentioned before, it is not always possible to calculate the pgf $S(z)$ of the buffer contents explicitly. Nevertheless, a technique has been developed to derive results concerning the moments and the tail distribution of the buffer contents from the associated functional equation. The technique involves considering those values for which the first argument(s) of the $P(., z)$ functions in both sides of the functional equation become equal (see e.g. [7], [68], [64], [66]).

## 2.4. Numerical example

As an illustration, we consider a statistical multiplexer to which messages consisting of a variable number of fixed-length packets arrive at the rate of one packet per slot (" train arrivals"), which results in a *primary* correlation in the packet arrival process. The arrival process contains an additional *secondary* correlation, which results from the fact that the distribution of the number of leading packet arrivals (of new messages) in a slot depends on some environment variable. This environment has two possible states 'A' and 'B', each with geometrically distributed sojourn times [47]. We compare the results obtained for this correlated train arrivals model with the results that would be found if a model without secondary correlation or an uncorrelated model for the packet arrival process were used. In Figure 1, the mean buffer contents for the three considered arrival models, i.e., $E[s]$ (correlated train arrivals), $E[s_{\text{prim}}]$ (uncorrelated train arrivals) and $E[s_{\text{un}}]$ (uncorrelated packet arrivals) are plotted versus the total load $\rho$, for different values of the environment correlation factor $K$, which can be seen as a measure for the absolute lengths of the sojourn times, when their relative lengths are

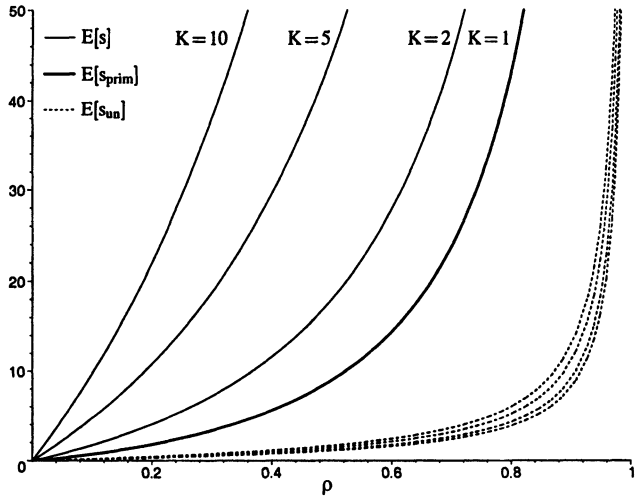


*Figure 1.* Mean buffer contents versus the total load $\rho$ for various values of $K$.

given [47]. The message-length distribution is a mixture of two geometrics according to the pgf $L(z) = \frac{0.5(1-\lambda)z}{1-\lambda z} + \frac{0.5(1-\tau)z}{1-\tau z}$ with mean 5 and a variance of 50. In an 'A'-slot, the number of new messages has a geometric distribution with mean 2, while no new messages are generated during 'B'-slots. The figure clearly shows the severe underestimation of the buffer contents when the different levels of correlation in the arrival process are neglected. Note that all the curves for $E[s_{\text{prim}}]$ coincide with the one representing $E[s]$ for $K = 1$ (uncorrelated environment). In the case of uncorrelated packet arrivals, $E[s_{\text{un}}]$ slightly increases with higher values of $K$, although not in the same drastic way as $E[s]$ in case of correlated train arrivals.

## 3. Performance Modeling using Matrix Analytic Methods

Two recent, and promising, developments within the area of matrix analytic methods are discussed in this section. It concerns the Markovian arrival process with marked arrivals, i.e., the MMAP[K] arrival process, and tree structured Markov chains of the $M/G/1$, $GI/M/1$ and Quasi-Birth-Death (QBD) type. While presenting these new developments, we mainly focus on their applicability towards telecommunication systems.

Matrix analytic methods, for queueing theory, found their origin in the 1960s in the work of Cinlar and Neuts [17]. During the 1970s, Neuts made a number of crucial contributions to the $M/G/1$ and $GI/M/1$ structures and wrote a book, the use of which is still widespread nowa-

days, on this subject [45]. During the 1980s, Neuts pursued his work at the University of Delawaire together with his associates and students Chakravarthy, Kumar, Latouche, Lucantoni and Ramaswami. In 1989, a second, perhaps somewhat less accessible to computer scientists, book [46] appeared on the $M/G/1$ structure that summarizes their achievements, it reflects the fact that the area of matrix analytic methods was growing vigorously. The theory and its applications have grown unabated ever since. This was clearly demonstrated in 1996, when the first conference on matrix analytic methods, and its applications, was organized. At the time of this writing a fourth conference will be held in July 2002 in Adelaide, Australia.

## 3.1. Markovian Arrival Process with Marked Arrivals

The usefulness of queueing theory as a means of analyzing the performance of telecommunication systems has been demonstrated extensively. However, until recently, most of the work done in this area applied to queueing systems that do not distinguish between customers, that is, all customers are of the same type and require the same type of service. There are plenty of applications were it would be suitable to distinguish between multiple customer types. For example, suppose that packets originating from $K$ different, possibly correlated, traffic streams form the input to a buffer, then it is often useful if we could obtain statistics, e.g., the delay distribution, for each individual source. The Markovian arrival process with marked arrivals, i.e., the MMAP[K] process, is an important building block that allows us to obtain such information.

Both a continues and a discrete time version of the MMAP[K] arrival process has been introduced [31, 27], but we restrict ourselves to the discrete time variant. We shall distinguish between two types of MMAP[K] processes: those that allow for batch arrivals to occur, and those that do not.

**MMAP[K] Process without Batch Arrivals.** A discrete time MMAP[K] arrival process that does not allow for batches to occur is a natural extension of the D-MAP arrival process [5]. Customers are distinguished into $K$ different types. The MMAP[K] is characterized by a set of $m \times m$ matrices $\{D_k \mid 0 \leq k \leq K\}$, with $m$ a positive integer. The $(j_1, j_2)^{th}$ entry of the matrix $D_k$, for $k > 0$, represents the probability that a type $k$ customer arrives and the underlying Markov chain (MC) makes a transition from state $j_1$ to state $j_2$. The matrix $D_0$

covers the case when there are no arrivals. The matrix $D$, defined as

$$D = \sum_{k=0}^{K} D_k,$$

represents the stochastic $m \times m$ transition matrix of the underlying MC of the arrival process. Let $\theta$ be the stationary probability vector of $D$, that is, $\theta D = \theta$ and $\theta e = 1$, where $e$ is a column vector with all entries equal to one. The stationary arrival rate of type $k$ customers is given by $\lambda_k = \theta D_k e$. Queues with MMAP[K] arrival processes are discussed in Section 3.1

**Example 3.1.** Consider a D-MAP arrival process characterized by the $m \times m$ matrices $\tilde{C}$ and $\tilde{D}$. Suppose that we wish to mark the arrivals by the state of the underlying MC at its generation time. This results in a MMAP[K] arrival process with $D_0 = \tilde{C}$ and with the matrices $D_k$, for $1 \leq k \leq K = m$, equal to zero, except for their $k$-th row, which is identical to the $k$-th row of $\tilde{D}$. Notice, the number of customer types might be smaller than $m$, because some rows of $\tilde{D}$ might be equal to zero.

**MMAP[K] Process with Batch Arrivals.** A discrete time MMAP[K] arrival process that allows for batches to occur —a natural extension of the D-BMAP arrival process [5]—is characterized by a set of $m \times m$ matrices $D_C$ where $C$ is a string of integers between 1 and $K$, that is, $C = c_1 \ldots c_b$ with $1 \leq c_l \leq K$ and $1 \leq l \leq b$. Let $b_{max}$ be the maximum batch size of the MMAP[K] arrival process. Let $\emptyset$ denote the empty string and $|C|$ the length of the string $C$. The $(j_1, j_2)^{th}$ entry of the matrix $D_C$, with $C \neq \emptyset$, represents the probability that a batch of $|C|$ arrivals occurs, while the underlying MC makes a transition from state $j_1$ to state $j_2$. The type of the $l$-th customer of the batch is $c_l$, for $1 \leq l \leq |C|$, if $C = c_1 \ldots c_{|C|}$. As before, $D = \sum_C D_C$ represent the transition matrix of the underlying MC and $\theta$ its stationary probability vector. The stationary arrival rate of type $k$ customers is given by $\lambda_k = \theta \sum_C N(C,k) D_C e$, where $N(C,k)$ counts the number of occurrences of the integer $k$ in the string $C$. Queues with MMAP[K] arrival processes are discussed in Section 3.1

**Example 3.2.** It is well known that a superposition of two, or more, D-BMAPs is again a D-BMAP. However, when superposing D-BMAPs customers generally loose their identity, meaning that we no longer know whether the arrival came from the first or the second D-BMAP. A MMAP[K] arrival process that eliminates this drawback can be con-

structed in the following way. Suppose that the first, resp. second, D-BMAP is characterized by the $m_1 \times m_1$ matrices $\tilde{D}_n^1$, resp. $m_2 \times m_2$ matrices $\tilde{D}_n^2$, for $n \geq 0$. Let $D_C$, with $C$ a string of $b_1 \geq 0$ ones followed by $b_2 \geq 0$ twos[1], be $m_1 m_2 \times m_1 m_2$ matrices. Instead of labeling the $m_1 m_2$ states $j$ of the underlying MC as 1 to $m_1 m_1$, we denote them as $(j, j')$, with $1 \leq j \leq m_1$ and $1 \leq j' \leq m_2$. The $(\mathbf{j_1}, \mathbf{j_2})^{th}$ entry, with $\mathbf{j_1} = (j_1, j_1')$ and $\mathbf{j_2} = (j_2, j_2')$, of the matrix $D_C$, with $C$ a string of $b_1$ ones followed by $b_2$ twos, equals $(\tilde{D}_{b_1}^1)_{j_1,j_2}\ (\tilde{D}_{b_2}^2)_{j_1',j_2'}$. A variety of examples is presented in [31, 29].

**The MMAP[K]/PH[K]/1 Queue.** In this section we discuss the MMAP[K]/PH[K]/1 queue with a first-come-first-serve (FCFS) and a last-come-first-serve (LCFS) service discipline. The service times of type $k$ customers, in a MMAP[K]/PH[K]/1 queue, have a common phase-type distribution function with a matrix representation $(m_k, \alpha_k, T_k)$, where $m_k$ is a positive integer, $\alpha_k$ is an $1 \times m_k$ nonnegative stochastic vector and $T_k$ is an $m_k \times m_k$ substochastic matrix. Let $T_k^0 = e - T_k e$, then the mean service time of a type $k$ customer equals $1/\mu_k = \alpha_k (I - T_k)^{-1} e$. The $i$-th entry of $\alpha_k$ represents the probability that a type $k$ customer starts its service in phase $i$. The $i$-th entry of $T_k^0$, on the other hand, represents the probability that a type $k$ customer completes its service provided that the service process is in phase $i$, while the $(i, j)$-th entry of $T_k$ equals the probability that it does not complete its service and the phase at the next time instance is $j$.

The positive recurrence, i.e., stability, of these queues was studied by He in [28]. Explicit formulas for the Laplace-Stieltjes transforms of the waiting times of a type $k$ customer have been obtained for a server with a FCFS service discipline [29]. An algorithm to obtain the steady state probabilities of a MMAP[K]/PH[K]/1 queue, where the MMAP[K] arrival process does not allow for batches to occur and the server follows a LCFS service discipline, is found in [30]. Finally, a simple algorithm, based on the $GI/M/1$ structure, has been developed to calculate the delay distribution of a type $k$ customer in a FCFS MMAP[K]/PH[K]/1 queue [54]. This algorithm is highly efficient if the MMAP[K] arrival process does not allow for large batch arrivals to occur.

**Example 3.3.** Let us continue with the MMAP[2] arrival process introduced in Example 3.2. Now, assume that each of the two D-BMAPs model a traffic source and that the traffic generated by both sources

[1]For simplicity, we assume that the arrivals of the first D-BMAP occur before those of the second, there is however no need to do so.

share a buffer. Moreover, assume that the packets generated by source $k$, for $k = 1, 2$, have a fixed length of $L_k$ bytes. Then, this buffer can be modeled by a discrete time MMAP[2]/PH[2]/1 queue, because fixed length service times have a phase type distribution. As a result, we could calculate the delay distribution of a source $k$ arrival using [54].

**Example 3.4.** Many random access algorithms (RAAs) that use grouped access as their channel access protocol (CAP) can be modeled in a natural way by means of a MMAP[K]/PH[K]/1 queue ([52, 55, 56]). When modeling such a RAA, a type $k$ customer corresponds to a group of $k$ contenders and its service time distribution is the time necessary for each of the $k$ contenders to successfully transmit their packet, starting from the completion time of the previous group.

## 3.2. Tree Structured Markov Chains

Another promising development in the theory of matrix analytic methods are tree structured Markov chains (MCs). Consider a discrete time bivariate MC $\{(X_t, N_t), t \geq 0\}$ in which the values of $X_t$ are represented by nodes of a $d$-ary tree, and where $N_t$ takes integer values between 1 and $m$. $X_t$ is referred to as the node and $N_t$ as the auxiliary variable of the MC at time $t$. A $d$-ary tree is a tree for which each node has $d$ children. The root node is denoted as $\emptyset$. The remaining nodes are denoted as strings of integers, with each integer between 1 and $d$. For instance, the $k$-th child of the root node is represented by $k$, the $l$-th child of the node $k$ is represented by $kl$, and so on. Throughout this paper we use lower case letters to represent integers and upper case letters to represent strings of integers when referring to nodes of the tree. We use '+' to denote concatenation on the right, e.g., if $J = j_1\ j_2\ j_3, k = j$ then $J + k = j_1\ j_2\ j_3\ j$. If $J$ can be written as $K_1 + K_2$ for some strings $K_1$ and $K_2$, $K_1$ is called *an ancestor* of $J$.

Algorithms that allow for the calculation of the steady state probabilities, have been identified for three subsets of the tree structured MCs, each subset allows for a certain type of transitions to occur:

- The *skip-free to the left*, i.e., M/G/1 Type, MCs: It is impossible to move from node $J$ to $\emptyset$, without visiting *all ancestors* of $J$ [51].
- The *skip-free to the right*, i.e., GI/M/1 Type, MCs: Transitions from a node $J$ are allowed to the root node $\emptyset$, the *children* of $J$ and the *children of all ancestors* of $J$ [70].
- The Quasi-Birth-Death (QBD) MCs: The chain can only make transitions to its parent, children of its parent, or to its children [69].

So far, the last subset has proven to be the most fruitful. Therefore, they are discussed in more detail in this section. If a tree structured QBD MC is in state $(J+k,i)$ at time $t$ then the state at time $t+1$ is determined as follows:

1 $(J,j)$ with probability $d_k^{i,j}, k = 1,\ldots,d$,

2 $(J+s,j)$ with probability $a_{k,s}^{i,j}, k,s = 1,\ldots,d$,

3 $(J+ks,j)$ with probability $u_s^{i,j}, s = 1,\ldots,d$.

Define $m \times m$ matrices $D_k, A_{k,s}$ and $U_s$ with respective $(i,j)^{th}$ elements given by $d_k^{i,j}, a_{k,s}^{i,j}$ and $u_s^{i,j}$. Notice that transitions from state $(J+k,i)$ do not dependent upon $J$, moreover, transitions to state $(J+ks,j)$ are also independent of $k$. When the Markov chain is in the root state $(J=\emptyset)$ at time $t$ then the state at time $t+1$ is determined as follows:

1 $(\emptyset,j)$ with probability $f^{i,j}$,

2 $(k,j)$ with probability $u_k^{i,j}, k = 1,\ldots,d$.

Define the $m \times m$ matrix $F$ with corresponding $(i,j)^{th}$ element given by $f^{i,j}$. Algorithms that calculate the steady state probabilities using the matrices $D_k$, $A_{k,s}$, $U_s$ and $F$ as input parameters are available in [69, 4].

**Example 3.5.** MMAP[K]/PH[K]/1 queue, where the MMAP[K] arrival process does not allow for batches to occur, with a last-come-first-serve (LCFS) service discipline can be modeled using a tree stuctured QBD MC [30]. Indeed, the line of customers waiting in a MMAP[K]/PH[K]/1 queue can be represented by a string of integers between 1 and $K$, thus as nodes of a $K$-ary tree. The auxiliary variable is used to represent the phase of the server, the type of customer in the server and the state of the MMAP[K] arrival process. The root node $\emptyset$ corresponds to a queue with a busy server and an empty waiting room. Therefore, one needs a generalized boundary condition to represent the situation where the waiting room is empty and the server is not busy. Information on generalized boundary conditions and other extension, i.e., MCs with a *forrest* structure, can be found in [70].

**Example 3.6.** Random access algorithms (RAAs) known as stack algorithms, or tree algorithms with free access, can be modeled using a tree structured QBD MC [53, 57]. As a result, it is possible to study the maximum stable throughput, as well as the mean delay, for various D-BMAP (and BMAP) arrival processes.

# 4. Asymptotic approximations for the performance evaluation of large broadband networks

## 4.1. The need for asymptotic methods

After a period of intensive development, multiservice broadband networks are now a reality. Current implementations already serve as high-speed backbone infrastructures and more extensive usage, accompanied by a further exploitation of these networks' advanced capabilities, is expected when the need for providing complex information services with strict quality guarantees will grow.

There are two primary performance-related characteristics that distinguish multiservice broadband networks from their "conventional" counterparts. The first is that, due to both the high transmission speed and the need for providing individualized—and stringent—quality of service (QoS) guarantees, very rare events (e.g., buffer overflows occurring with probability as low as $10^{-6}$, or smaller) become significant. Consequently, most relevant performance metrics must be based on distribution tails rather than mean values. The second characteristic is that most bandwidth-demanding traffic types appearing on broadband networks are bursty, i.e., they feature significant rate excitations and correlated packet interarrival times. These are properties that leave a mark on the queueing phenomena governing the network's performance.

The typical queueing effects of burstiness are demonstrated by the main graph of Fig. 2, depicting the buffer overflow probability (a standard performance metric) at a network multiplexer or switch loaded by a superposition of bursty traffic streams, as a function of the buffer size. Two distinct regions are clearly identified: In the first region (small buffer sizes) the rate correlations do not become apparent, the traffic is primarily characterized (at the, so called, 'packet level') by properties of the individual interarrival times between successive packets, and the overflow probability decays rapidly with increasing buffer size (at an exponentially fast rate, since the graph uses a log-linear scale). In the second region (larger buffer sizes) the rate correlation details (usually collectively called 'burst level traffic properties') become noticeable, resulting in a quite smaller rate of decay for the overflow probability.

Clearly, accurate prediction of tail probabilities—like those in the example—requires the usage of sophisticated traffic models, able of providing a sufficiently precise characterization of traffic at both the packet and burst levels. Such detailed models, and associated analysis methods, do exist and are invaluable whenever thorough queueing analysis is called for. In due account, the paper reviews two important classes

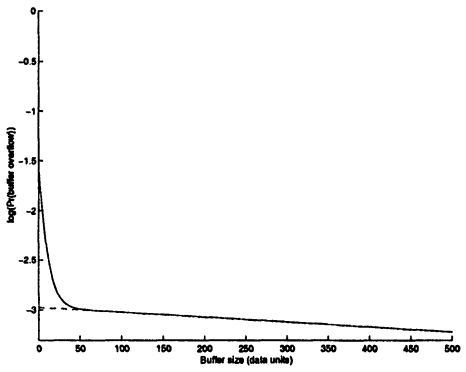


*Figure 2.* The effect of burstiness: overflow probability vs buffer size (at a log-linear scale).

of models/solution methods (see the sections on matrix analytic techniques, and on generating functions based techniques). Unfortunately, detailed descriptions suffer from the 'state space explosion' problem. Indeed, the state spaces of models for all but the simplest traffic patterns have to be rather large, if both the packet- and burst-level behavior is to be captured. The situation becomes worse when it is realized that, in virtually all congestion phenomena of interest, the aggregate traffic load consists of a (frequently heterogeneous) superposition of a large number of individual streams and that the state space of the model for the aggregate traffic depends factorially on the—already large—spaces of the constituents.

In an attempt to partially alleviate this difficulty, 'fluid-flow' models of traffic have been proposed. These models disregard the discrete nature of the packet level details, representing traffic as the flow of a continuous fluid (hence their name). The instantaneous rate of this flow is taken equal to the average rate of the real traffic over an appropriate time window, large enough to "hide" the packet details, but also small enough to preserve the burst-level rate fluctuations. This approach has been quite successfully employed towards the accurate representation of burst-level traffic dynamics with a reduced set of model parameters. An example is provided by the dashed graph in Fig. 2, which represents the overflow probability curve corresponding to the fluid-flow counterpart of the original traffic and which matches quite satisfactorily with the exact result over the burst level region. For further information on (primarily Markovian) fluid-flow models see: [1, 40, 43, 49] for the basic theoretical foundation and analysis techniques, [3, 37, 42] for embelishments of the theory and efficient computational algorithms, and [38, 44] for

multiple-scale phenomena occurring when the traffic possesses burst-level dynamics with a finer structure.

However, although the fluid-flow concept works for reducing the complexity of models for *individual* traffic streams, it cannot alleviate the state space explosion due to superposition. For this reason, many important performance-related network mechanisms, particularly those that must operate within a short time-frame (such as on-line traffic control) or over a combinatorially large domain (e.g., network-wide resource (re)allocation), cannot rely on "classical" queueing techniques, even the fluid-flow ones.

Fortunately, there's still a viable way of addressing the problem, grounded on the fact that modern broadband networks are, in some respects, "large" systems, featuring high link capacities and large switches, and requiring that probabilities of hazardous events (like overflows leading to data losses) be very small (so as to provide reliable QoS guarantees). This setting suits well to the 'Theory of Large Deviations' (TLD), a body of theoretical results and techniques that address systems "scaled up" by a large parameter and examine the circumstances under which associated (scaled) random variables may attain values in a designated set with an exponentially small probability, asymptotically as the scaling parameter approaches infinity. TLD may be used to compute the rate of exponential decay in the probabilities of interest and, moreover, determine the way in which these 'rare events' occur. A comprehensive general treatment of TLD can be found in, e.g., [16], while [10] provides a less formal exposition, explicitly geared towards applications. Reference [63] may be consulted for a brief overview of topics and further references.

Building on the TLD foundations, the very same characteristics that lead to state-space explosion in "conventional" models have been exploited towards the development of asymptotic theories that quantify congestion in broadband networks under bursty load. The purpose of this section is to give an outline of the relevant results. Before embarking on the review, however, it is important to note that, besides analytical tractability, a prime advantage of the asymptotic methods is their potential for conceptual clarity, something crucial for highlighting the effect of fundamental phenomena in explicit terms.

Generically, two such congestion-related phenomena may be identified: The first, frequently called 'multiplexing gain', relates to the fact that (as a consequence of the law of large numbers) aggregation of many independent traffic streams results in smoother compound traffic, reducing the probability with which the aggregate data rate raises above its mean value. As more streams are multiplexed, the amount of bandwidth

per stream required to compensate for the rate excitations is reduced (for a given QoS requirement), justifying the name of the phenomenon. In the absence of significant buffering, multiplexing gain is the only mechanism through which QoS may be attained while using less bandwidth than peak-rate. In Fig. 2 this is reflected at the non-negligible probability of overflow even with a zero buffer size. The relevant asymptotic theory is reviewed in Subsection 4.2.

The second fundamental phenomenon relates to another mechanism of controlling rate excitations so as to avoid data losses, that of temporarily storing excessive data into a buffer. The larger the buffering resource, the smaller the capacity requirement for the output port becomes, for a given loss probability. In analogy with multiplexing gain, this bandwidth-savings effect will be called 'buffering gain'. In Fig. 2 it is reflected at the decay of the overflow probability with increasing buffer size, even at the "slow" burst-level region. The asymptotic theory relevant to buffering gain is reviewed in Subsection 4.3.

The two regimes just outlined relate to either no buffer, or a large buffer, so that either the multiplexing gain, or the buffering gain dominate, respectively. In many cases the available buffer is neither negligible nor dominant and both phenomena are noticeable. For this more general setting there is an improved asymptotic theory that can quantify the combined effect of both gain factors, by considering systems where the load and resources (buffer and bandwidth) are proportionally scaled by a large parameter. Elements of this theory are provided in Subsection 4.4.

## 4.2. Asymptotics for multiplexers with small buffers

Consider a multiplexer (or an output port unit of a switch) featuring a negligibly small buffer and serving traffic through an output link of capacity equal to $C$. The aggregate traffic loading this system can be described as a stochastic instantaneous-rate[2] process $\{r(t),\ t \in \mathbb{R}\}$, which it is assumed throughout stationary. Tracking just instantaneous rates is adequate, as there is no buffer to "record the past history" of the traffic. In the following, the properties of the instantaneous rate will be described through the respective log-moment generator (also called the 'cumulant generator') $\phi(s) \widehat{=} \log \mathbb{E}\, e^{sr(t)}$. As an implication of stationarity, $\phi(s)$ is independent of time.

[2]Here we adopt a fluid approach and represent the flow of data as a continuum. However, all results in this section bear obvious analogies with a discrete-time setting, in which $r(t)$ stands for the amount of data contributed during the time-slot indexed by (the now integer) $t$.

At this point it is reminded that the log-moment generator of a random variable (r.v.) is a convex function (actually strictly convex, unless the r.v. is a.s. constant). The set $\{ s \in \mathbb{R} \mid \phi(s) < \infty \}$ is called the generator's 'effective domain'. If $s = 0$ is in the interior of this domain (a mild condition, assumed throughout and satisfied in all cases of practical interest, in particular when the r.v. is bounded—translated to the existence of a finite peak rate in our case), then the generator is an analytic function on the whole interior of its effective domain. By convexity, the derivative $\phi'(s)$ is increasing (strictly increasing if the r.v. is not a.s. constant) and the same may be shown for $\phi(s)/s$. Furthermore, the limits of these functions are related to the extremal values[3] of the corresponding r.v. $X$ as follows:

$$\begin{aligned} \operatorname{ess\,inf} X &= \lim_{s\to-\infty} \phi'(s) = \lim_{s\to-\infty} \frac{\phi(s)}{s} < \lim_{s\to 0} \frac{\phi(s)}{s} = \mathbb{E}\, X \\ &= \lim_{s\to 0} \phi'(s) < \lim_{s\to+\infty} \frac{\phi(s)}{s} = \lim_{s\to+\infty} \phi'(s) = \operatorname{ess\,sup} X. \end{aligned} \tag{12}$$

Since there is no buffer, overflows (and data losses) occur whenever the instantaneous data rate exceeds the system's capacity. We now derive an upper bound to the probability of overflow. Indeed, by a Chebycheff-type argument, for any $s \geq 0$,

$$\begin{aligned} \Pr\{ r(t) > C \} = \int_{x=C^+}^{\infty} dF_r(x) &\leq \int_{x=C^+}^{\infty} e^{s(x-C)}\, dF_r(x) \\ &\leq \int_{x=0}^{\infty} e^{s(x-C)}\, dF_r(x) = \exp\{\phi(s) - sC\}. \end{aligned}$$

By taking logarithms and optimizing over the permissible range of parameters, we obtain

$$\log \Pr\{ r(t) > C \} \leq -\sup_{s\geq 0}\{Cs - \phi(s)\}. \tag{13}$$

This bound is known in the literature as 'Chernoff's bound'. Assuming a stable system (i.e., $C > \mathbb{E}\, r(t)$), the maximum over nonnegative reals coincides with the maximum over the entire real line, i.e.,

$$\forall C > \mathbb{E}\, r(t) \mathrel{\widehat{=}} \bar{r}, \qquad \sup_{s\geq 0}\{Cs - \phi(s)\} = \sup_{s\in\mathbb{R}}\{Cs - \phi(s)\} \mathrel{\widehat{=}} I(C), \tag{14}$$

[3]The upper extremal value of a r.v. $X$, called 'essential supremum' and denoted by $\operatorname{ess\,sup} X$, is the largest value that $X$ is not improbable to exceed, namely $\operatorname{ess\,sup} X = \sup\{ x \in \mathbb{R} \mid \Pr\{X > x\} > 0 \}$. The lower extremum, called 'essential infimum' and denoted by $\operatorname{ess\,inf} X$, is defined analogously.

the value of the Fenchel-Legendre transform of $\phi(\cdot)$ at $C$. Furthermore, it may be shown that, for $C > \bar{r}$, the Fenchel-Legendre transform $I(\cdot)$ is an increasing function (actually strictly increasing, unless $r(t)$ is a.s. constant), expressing the intuitively appealing fact that the overflow probability becomes smaller as the system's capacity increases.

Assume now that the aggregate traffic consists of a large number $n$ of independent and identically distributed streams, while the system's capacity is proportionally scaled, maintaining a fixed amount of bandwidth per source, i.e., $C = nc$. Since log-moment generators are additive for independent r.vs, the aggregate generator is $\phi_n(s) = n\phi(s)$ (where now $\phi(\cdot)$ signifies the generator of a single stream) and from equations (13) and (14) it follows that the overflow probability is bounded by $e^{-nI(c)}$, decaying exponentially with large $n$ at a rate equal to $I(c)$. This reflects the fact that, as more sources are multiplexed and the bandwidth per source $c$ remains fixed, overflows become less probable, because the compound traffic "smoothens". In other words, and due to the monotonicity of $I(\cdot)$, a smaller value of $c$ is required as $n$ increases, for a given target overflow probability. This is exactly the multiplexing gain phenomenon, discussed in the previous subsection.

The Chernoff bound of eq. (13) is conservative, allowing for safe performance-related decisions. Not only that, but the bound is asymptotically tight, as the number of sources $n \to \infty$. Specifically, by Cramér's Theorem (see, e.g., [16, Theorem 2.2.3]), it holds

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{ r_n(t) > nc \} = -I(c), \tag{15}$$

where, as with the generator, $r_n(t)$ denotes the aggregate rate. This result suggests that, when $n$ is large enough, the probability of overflow is $e^{-\epsilon}$, where $\epsilon = nI(c) + o(n)$. (The quantity $\epsilon$ expresses the achievable QoS at a logarithmic scale and will be called the 'quality level' in the sequel.) There is also a more detailed result, called the 'Bahadur-Rao' correction, that strengthens the asymptotic equivalence of (15) to linear, rather than logarithmic order. (In this result, $I(c)$ still remains the dominant factor determining the probability of overflow.) For details, see, e.g., [16, Theorem 3.7.4].

When the traffic is a heterogeneous mix of independent traffic streams, the previous theory still applies. Indeed, consider $k$ traffic classes, each containing $n_i$, $i = 1, \ldots, k$ independent and identical streams. Then the total number of sources is $n = \sum_{i=1}^{k} n_i$ and the aggregate generator is constructed by the individual counterparts through $\phi_n(s) = \sum_{i=1}^{k} n_i \phi_i(s)$. In this setting (15) still holds, i.e., for large $n$ the prob-

ability of overflow is approximately $e^{-\epsilon}$, with quality level $\epsilon = nI(c) = \sup_s\{Cs - \sum_{i=1}^{k} n_i\phi_i(s)\}$.

We now discuss the computation of the decay rate in the asymptotic (15). Due to the convexity of log-moment generators, the function to be maximized in (14) is concave and attains a unique maximum. Moreover, by differentiability (again borrowed from the generator) the derivative of the function in (14) is zero at the maximizing argument. From these observations and from eq. (12) it follows that when the capacity $C$ is between the aggregate mean and peak rates, the quality level is computed as

$$\epsilon = nI(C/n) = \sup_{s\geq 0}\Big\{Cs - \sum_{i=1}^{k} n_i\phi_i(s)\Big\} = Cs^* - \sum_{i=1}^{k} n_i\phi_i(s^*), \qquad (16)$$

where $s^*$ is the unique[4] argument satisfying

$$\sum_{i=1}^{k} n_i\phi_i'(s^*) = C, \qquad (17)$$

and where the equations have been expressed in a form suitable for a general heterogeneous traffic mix.

Usually, (16) and (17) must be solved numerically. However, the canonical example of a homogeneous on/off traffic mix admits a closed form solution. Indeed, for any on and off sojourn distributions (just assuming finite means, respectively $\mathbb{E}\,T_{\text{on}}$ and $\mathbb{E}\,T_{\text{off}}$) each constituent rate process is stationary and ergodic. By letting $p = \mathbb{E}\,T_{\text{on}}/(\mathbb{E}\,T_{\text{on}} + \mathbb{E}\,T_{\text{off}})$ stand for the probability of visiting the on-state, the instantaneous rate of a single stream is Bernoulli distributed, with generator $\phi(s) = \log[pe^{sr} + (1-p)]$, where $r$ is the stream's peak rate. Then, application of (17) and (16), yields

$$\epsilon = n\left(\beta\log\frac{\beta}{p} + (1-\beta)\log\frac{1-\beta}{1-p}\right), \qquad \text{where} \quad p < \beta \mathrel{\widehat{=}} \frac{C}{nr} < 1.$$

Up to this point, the focus of the discussion was on estimating the system's performance under given resources and traffic load. However, network traffic engineering usually deals with problems of an "inverse" nature. One particularly important one is the so called, traffic admission control (also named connection admission control—CAC), where the network resources (in our case the multiplexer's capacity $C$) and the

[4]Except in the trivial case where all traffic rates are a.s. constant. This case is excluded here.

desired quality level $\epsilon$ are given and the task consists of deciding whether a candidate traffic mix may be admitted by the network while still satisfying the QoS requirement. Formally, assume that the traffic load at a multiplexer may consist of a superposition of streams from $k$ different traffic classes, each with known characteristics (quantified through the respective generators $\phi_i(\cdot)$, $i = 1, \ldots, k$) and let a potential traffic mix be represented by the vector $\boldsymbol{n} = (n_1, \ldots, n_k)$, with elements the numbers of streams from each class participating in the mix. In this notation, a traffic mix may be admitted without violating the QoS, iff it belongs to the so called admission domain $\{ \boldsymbol{n} \mid f(\boldsymbol{n}) \geq \epsilon \}$, where $f(\boldsymbol{n})$ stands for the right hand side of (16).

Given this framework, traffic admission control could in principle proceed by computing $f(\boldsymbol{n})$ through (17) and (16) and comparing the result to the target quality level $\epsilon$. However, the relevant computations involve *all* traffic classes in the mix, making it difficult to take incremental decisions (useful in the common case when a single new connection asks to join an already accepted—potentially large—mix). For this reason, alternative algorithms are required, which usually rely on determining the boundary of the admission domain (i.e., mixes satisfying $f(\boldsymbol{n}) = \epsilon$). If that boundary was linear, then a particularly simple algorithm would be possible, because there would be constants $a_i$, $i = 1, \ldots, k$ and $b$ (possibly dependent on $C$ and $\epsilon$ but not on the traffic mix), such that the admission domain would contain exactly those $\boldsymbol{n}$ satisfying

$$\sum_{i=1}^{k} a_i n_i \leq b. \tag{18}$$

Thus, for the purposes of admission control, each traffic stream would be completely characterized by the quantity $a_i$ corresponding to its class and incremental admission control would proceed by merely adding this quantity to a register (maintaining the sum for the already present traffic) and comparing to $b$.

Unfortunately, the boundary of the admission domain, as defined by (16) and (17) is not linear[5]; the typical form of its shape is displayed on Fig. 3 for two traffic classes (ignore for the moment the linear segment). Despite this difficulty, it is still possible to obtain a locally optimal linearization, by observing that, due to (16), $f(\boldsymbol{n})$ is convex and the same holds for the complement of the admission domain. Thus, it is assured that the tangent hyperplane at a point $\boldsymbol{n}^*$ on the boundary will

[5]Nonlinearity is unavoidable if the nature of the multiplexing gain phenomenon is to be preserved. This point will be discussed to a greater extent in Subsection 4.3.

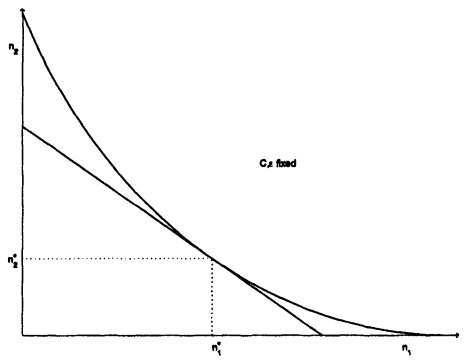

*Figure 3.* Admission domain for two traffic classes and linear approximation of the boundary around $\boldsymbol{n}^*$

rest inside the admission domain (see Fig. 3), while also coinciding with the true boundary at $\boldsymbol{n}^*$. By observing that $f(\boldsymbol{n}^*) = \epsilon$ and by using (16) and (17), it follows that $\partial f/\partial n_i|_{\boldsymbol{n}=\boldsymbol{n}^*} = -\phi_i(s^*(\boldsymbol{n}^*))$ and, further, that the subset of the domain bounded by the tangent hyperplane contains those traffic mixes $\boldsymbol{n}$ satisfying

$$\sum_{i=1}^{k} n_i \frac{\phi_i(s^*(\boldsymbol{n}^*))}{s(\boldsymbol{n}^*)} \leq C - \frac{\epsilon}{s^*(\boldsymbol{n}^*)}. \tag{19}$$

In order to use (19), one must determine a traffic mix $\boldsymbol{n}^*$ at the boundary of the true admission domain and then compute the corresponding value of the maximizing $s$-parameter, namely $s^*(\boldsymbol{n}^*)$. Although these initialization steps require rather heavy computations, the actual admission control through (19) is simple, because the latter is of the simple form (18). However, note that, since the linearization is optimal only with respect to the chosen $\boldsymbol{n}^*$, successive connection admissions (and terminations) may move the current traffic mix away from the initial choice $\boldsymbol{n}^*$, at a vicinity of the domain for which the linearization is overly conservative (see the figure), thus resulting in a waste of network resources. In such a case, a new boundary point close to the current traffic mix should be chosen and the linearization procedure around it should be applied afresh.

We close this subsection by noting that, while the basic asymptotic performance estimate is a standard result in the Theory of Large Deviations (and thus known for many years), its application in the study of broadband networks and, in particular, the results on admission domains

and the linearization of their boundaries were originally contributed by Hui [32, 33].

## 4.3. Asymptotics for large buffers: effective bandwidth theory

We now turn into the study of multiplexers that feature large buffering capabilities. Like previously, we seek to present a theory linking the traffic load and the network resources (viz., the amount of buffer memory and the output link's capacity) to the probability of data loss due to buffer overflow, the latter serving as the performance metric. While in the bufferless setting it was adequate to represent the traffic characteristics through instantaneous rate properties, this subsection deals with large buffers that expose the properties of rate correlations over large time intervals. Therefore, it is necessary to study random variables of the form $V(\tau, \tau + t)$, denoting the amount of data generated over the interval $(\tau, \tau + t]$. It will be assumed throughout that the data process has stationary increments[6], i.e., $V(\tau, \tau + t)$ depends only on the length $t$ of the time-interval, not its origin, and can be denoted simply as $V(t)$. By virtue of stationarity, $\mathbb{E}\,V(t) = \bar{r}t$ for all time-lengths $t$, $\bar{r}$ being the mean traffic rate. Further stochastic properties of $V(t)$ will be described through the corresponding log-moment generator

$$\phi(\theta, t) \,\widehat{=}\, \log \mathbb{E}\, e^{\theta V(t)}, \tag{20}$$

for which two relevant conditions are introduced:

**C1** For each $\theta$, the limit $\phi_\infty(\theta) = \lim_{t\to\infty} \frac{\phi(\theta,t)}{t}$ exists and is finite.

**C2** $\phi_\infty(\theta)$ is strictly convex and differentiable.

Condition C1 ensures that the traffic is not long-range dependent (a case for which the theory, in the form presented here, does not hold), while Condition C2 is a guarantee that the strict convexity and differentiability of the generator $\phi(\theta, t)$ will also be inherited by the limit.

Under Condition C1, the 'effective bandwidth function' (EBF) of the traffic is defined as

$$a(\theta) = \phi_\infty(\theta)/\theta, \qquad \theta \geq 0. \tag{21}$$

As a log-moment generator, $\phi(\theta, t)$ is convex in $\theta$, a property also transferred to the limit $\phi_\infty(\theta)$ as well. Thus, according to the discussion early

[6]This assumption holds in particular when data are generated according to a stationary rate process $\{r(t),\ t \in \mathbb{R}\}$, since in that case $V(\tau, \tau + t) = \int_\tau^{\tau+t} r(x)\,dx$.

in Subsection 4.2, the EBF $a(\cdot)$ is an increasing function. Furthermore, if Condition C2 also holds, then $a(\cdot)$ is strictly increasing. Lastly, observe that, by virtue of (12),

$$\bar{r} = \lim_{t\to\infty} \frac{\mathbb{E}\, V(t)}{t} = \phi'_\infty(0) = a(0) \le a(\theta) \le \lim_{\theta\to\infty} a(\theta) = \lim_{\theta\to\infty} \frac{\phi_\infty(\theta)}{\theta}$$
$$= \lim_{t\to\infty} \frac{\operatorname{ess\,sup} V(t)}{t} \triangleq \hat{r},$$

establishing that the EBF is bounded between mean and peak rate. (The peak rate $\hat{r}$ is with respect to an asymptotically large time-window and may, in some cases, be smaller than the instantaneous peak rate.)

The importance of the EBF is due to the following properties: Assume that traffic of EBF $a(\cdot)$ loads a multiplexer featuring infinite buffer space and an output link of capacity $C$. Further, assume there is some $\theta > 0$, such that $a(\theta) < C$. Then, it may be proved that the distribution tail of the queue content $Q(t)$ has at all times an exponential upper bound of rate $\theta$. In other words, there exists a constant $d(\theta)$, such that

$$\Pr\{\, Q(t) > B \,\} \le d(\theta) e^{-\theta B}, \qquad \forall t \ge 0,\ \forall B \ge 0.$$

There is also a "reciprocal" result: If $a(\theta) > C$ the capacity is not large enough and it may be shown that the distribution tail of the queue content cannot be bounded exponentially using rate $\theta$.

These two statements taken together suggest that, in order to achieve an exponential decay of at least rate $\theta$ for the overflow probability under increasing buffer size, the system's capacity must be greater than $a(\theta)$. In this case, the achievable decay rate is $\theta^* = \sup\{\, \theta \mid a(\theta) < C \}$. Obviously, when the EBF is strictly increasing (as when Condition C2 holds), $\theta^* = a^{-1}(C)$. In fact, for this case the following stronger assertion can be made: If, besides Condition C1, C2 also holds, the buffer content $Q(t)$ has a stationary distribution with tail satisfying

$$\lim_{B\to\infty} \frac{-\log \Pr\{\, Q > B \,\}}{B} = \theta, \qquad \text{where} \quad \theta = a^{-1}(C). \tag{22}$$

This result not only establishes asymptotic exponentiality for queue tails, but may also be used to determine the bandwidth requirements, as a function of the buffer size and the QoS level.

Indeed, assume that the multiplexer has a large (but finite) buffer size $B$ and set the requirement that the system overflows with probability at most $e^{-\epsilon}$. (This specifies a quality level equal to $\epsilon$ in the terminology of the previous subsection.) Then, by (22), one must ensure that $\theta \ge \epsilon/B$ or, equivalently, $C \ge a(\epsilon/B)$, which is the desired

result. Although this last relation is in a form suitable for admission control, it must be remembered that $a(\cdot)$ is the EBF for the *whole* traffic load, thus it depends on the properties of all multiplexed streams. Fortunately, the definition of the EBF by (21) and the additivity of log-moment generators over independent r.vs, ensure that, for a traffic mix $\boldsymbol{n} = (n_1, \ldots, n_k)$, containing $n_i$ streams of class $i$, for $i = 1, \ldots, k$, the aggregate EBF is simply $a(\theta) = \sum_i n_i a_i(\theta)$. In particular, the relation for the bandwidth requirements becomes

$$a(\epsilon/B) = \sum_{i=1}^{k} n_i a_i(\epsilon/B) \leq C, \tag{23}$$

specifying a linear boundary of the form (18) for the admission domain and enabling the particularly simple algorithm for incremental admission control discussed in Subsection 4.2.

As a matter of fact, the name 'effective bandwidth' is exactly due to the linearity in (23), as the quantity $a_i(\epsilon/B)$ determines, *independently of the rest of the traffic environment* the amount of bandwidth that must be granted to a source of class $i$, in order to satisfy the QoS requirements with the given amount of buffering. Due to this independence, each traffic stream behaves, in a sense, like a constant-rate counterpart; for this reason effective bandwidths are sometimes called 'effective rates' or 'equivalent bandwidths'. It is mentioned that originally the term was introduced by [32], in connection with (19). However, since the linearization in (19) is only locally significant, the term is now mostly used in the sense (23), for the large-buffer regime.

Note that the linearity precludes any potential for bandwidth savings due to multiplexing gain. Indeed $n$ traffic streams require bandwidth $C = na(\epsilon/B)$, thus maintaining a constant bandwidth per source $C/n$, no matter how large $n$ becomes. This is not surprising, as the theory holds asymptotically as the buffer size $B \to \infty$ when the multiplexing gain is negligible, compared to the buffering gain effect.

At this point it is remarked that the effective bandwidth theory was developed through a series of contributions. The asymptotic exponentiality of distribution tails for the stationary queue content and the implications for this on a linear admission domain were originally established for iid, Markovian on/off, and other simple traffic models [23, 25, 34] and were later generalized for the class of arbitrary Markovian fluids [22]. An extended theory that covers more general stationary rate processes followed [36, 11, 24], making explicit use of results from Large Deviations Theory. Furthermore, a modification [21] of the limiting generator $\phi_\infty(\theta)$, using a time scaling more general than linear, allowed

the treatment of traffic with long range dependence. See [12] for a review of the effective bandwidth theory along the statistical mechanics viewpoint and [58] for a discussion of resource management techniques based on the effective bandwidth concept. Further references may be found in [35].

Apart from the general properties discussed earlier, the particular form of the EBF $a(\cdot)$ depends on stochastic details specific to the corresponding traffic stream. To review some examples, consider Markovian on/off fluid models, featuring a peak rate $r$ and exponentially distributed on and off sojourns with mean durations $\tau$ and $\sigma$, respectively. In this case the EBF takes the form

$$a(\theta) = \frac{1}{2}\left(r - (\frac{1}{\tau} + \frac{1}{\sigma})\frac{1}{\theta} + \sqrt{\left(r - (\frac{1}{\tau} + \frac{1}{\sigma})\frac{1}{\theta}\right)^2 + \frac{4r}{\sigma\theta}}\right),$$

a result that originally appeared in [23] and was further exploited in [25]. In the more general case of arbitrary Markovian fluids, traffic is described through a 'rates-matrix' $R = \mathrm{diag}\{r_1, \ldots, r_n\}$ and the infinitesimal generator $M$ of a continuous-time Markov Chain, which governs the transitions between rate values. For this class of models it has been shown [22] that the EBF is $a(\theta) = \lambda_{\max}\left(R + \frac{1}{\theta}M\right)$, i.e., the largest eigenvalue of the essentially nonnegative matrix $R + \frac{1}{\theta}M$. A further generalization [39] allows the explicit calculation of effective bandwidths corresponding to semi-Markovian fluids, i.e., models where transitions between rates are still Markovian, but the periods during which rate values are sustained may be arbitrarily distributed (but not heavy-tailed). In this case, the EBF is determined through an implicit function problem, derived from the requirement that the spectral radius of an appropriate nonnegative matrix be equal to unity. For general on/off traffic streams, of peak rate $r$, this result simplifies as follows: Let $\phi_+(s)$ and $\phi_-(s)$ stand for the log-moment generators corresponding to the distributions of the on and off sojourns, respectively. Then, for any $\theta > 0$, the EBF is $a(\theta) = u(\theta)/\theta$, where $u(\theta)$ is the unique positive solution of

$$\phi_+(r\theta - u) + \phi_-(-u) = 0.$$

We close this subsection by mentioning that, instead of adopting a traffic model and trying to determine the EBF through it (something not always feasible), there are alternative approaches, which target the direct measurement of the EBF, thus bypassing modeling. For work along this line, see, e.g., [19, 13].

## 4.4. Scaling the system's size

The two previous asymptotic regimes were appropriate for either very large buffers or very small ones. However, there are cases where the buffering resource is neither negligible nor overly dominant and then both the multiplexing- and buffering-gain effects are noticeable and must be taken into account. We now briefly discuss results for this more general setting. The relevant asymptotic regime assumes a large number of traffic streams $n$ and proportionally scaled (large) buffer $B$ and bandwidth $C$. In other words, $B = bn$ and $C = cn$, maintaining a constant amount of resources per stream, as $n \to \infty$. This type of scaling was originally introduced by [62], in connection with traffic consisting of exponential on/off fluids.

In our setting, each traffic stream is a data generation process, which, as in Subsection 4.3, is assumed to have stationary increments. The generator (20) is again used as the traffic descriptor. (Generalizations, relaxing the assumption on stationarity or the requirement for iid streams exist.) Let the stationary queue content under a load of $n$ traffic streams be denoted as $Q_n$; then the probability of overflow is $\Pr\{ Q_n > bn \}$. The basic result [6, 14] (also [48], for the particular context of general on/off fluids) is that, under some regularity conditions, notably the validity of Condition C1 in Subsection 4.3,

$$\lim_{n\to\infty} \frac{-\log \Pr\{ Q_n > bn \}}{n} = I(c,b) \widehat{=} \inf_{t>0} \sup_{\theta} \{(ct+b)\theta - \phi(\theta,t)\}. \quad (24)$$

There is also a generalization [18] which, among other things, relaxes the requirement for Condition C1 (by introducing a different time-scaling for the generator) and is appropriate for usage with long range dependent traffic.

For a heuristic explanation of (24) remember that, by Lindley, $Q_n = \sup_{t>0}(V_n(t) - nct)$, where $V_n(t)$ is the total amount of data generated by the $n$ streams over time $t$. Then (24) is essentially Cramér's asymptotic on $V_n(t) - nct$ (see (15) and (14)), followed by an optimization of the time scale (using Laplace's principle of the 'dominating term'). Note that the appropriate time scale relevant to the result is neither $t = 0$ (as was the case with bufferless systems, where only instantaneous rates were needed) nor $t = \infty$ (which was appropriate for very large buffers), but actually the argument extremizing (24), say $t^*$. The value of $t^*$ depends on $c$ and $b$ and expresses the relative importance of the multiplexing- and buffering-gain. Indeed, by assuming differentiability, (24) implies that

$$t^* = \frac{\partial I(c,b)}{\partial c} \Big/ \frac{\partial I(c,b)}{\partial b} = -\left.\frac{\partial b}{\partial c}\right|_{I(c,b)=\epsilon}, \quad (25)$$

thus $t^*$ quantifies the (local) tradeoff between bandwidth and buffer for a given quality level per source. More specifically, it is possible to formally define a 'buffer-bandwidth' curve of the form $b(c, \epsilon)$, that describes the amount of buffering required for achieving a desired quality level, given the capacity (all quantities being scaled by the number of sources). It may be shown that this curve is convex [20, 41]. Since, by (25), $dc/db = -1/t^*$, the inverse curve is also convex and decreasing, implying that $t^*$ increases with $b$. This is intuitive, as larger buffers pronounce more the traffic correlations, requiring a larger time-scale for their representation.

Still in connection with this point, it has been shown [6, 14] that as $b \to \infty$ then $t^* \to \infty$ and $I(c, b)$ tends to the asymptotic (22). Similarly, when $b \to 0$, then also $t^* \to 0$ and $I(c, b)$ tends to the asymptotic of (15) and (14), where in place of the instantaneous rate generator the limit $\lim_{t \to 0} \phi(s/t, t)$ is implied[7]. Thus the theories for small and large buffers may be regarded as special cases of the results in this subsection. Note however, that application of (24) is considerably more difficult than the other asymptotic results, not only because the generator $\phi(\theta, t)$ must be determined for all time-scales instead of at a limiting value, but also because the minimization with respect to time is non-convex (unlike the maximization in $\theta$) and thus difficult to perform numerically.

We close by noting that it is possible to define an admission domain for the more general regime of this subsection. This domain is neither linear (as in Subsection 4.3) nor possessing a convex complement (as in Subsection 4.2). However, it is still possible to obtain a local linearization, around points on the boundary, thus introducing a (locally significant) notion of effective bandwidth for this case too. For more details see [35].

## References

[1] D. Anick, D. Mitra, and M. M. Sondhi. "Stochastic theory of a data-handling system with multiple sources". *Bell System Tech. J.*, 61(8):1871–1894, October 1982.

[2] Y. Xiong B. Steyaert and H. Bruneel. An efficient solution technique for discrete-time queues fed by heterogeneous traffic. *International Journal of Communication Systems*, 10:73–86, 1997.

[3] A. Baiocchi, N. Blefari-Melazzi, A. Roveri, and F. Salvatore. "Stochastic fluid analysis of an ATM multiplexer loaded with heterogeneous on-off sources: an effective computational approach". In *Proc. INFOCOM '92*, pages 3C.3.1–3C.3.10, 1992.

[7] This limit coincides with the definition in Subsection 4.2 for fluid traffic.

[4] D. A. Bini, G. Latouche, and B. Meini. Algorithms for tree-like stochastic processes. *In Preparation*, 2002.

[5] C. Blondia. A discrete-time batch markovian arrival process as B-ISDN traffic model. *Belgian Journal of Operations Research, Statistics and Computer Science*, 32(3,4), 1993.

[6] D. D. Botvich and N. G. Duffield. "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers". *Queueing Sys.*, 20:293–320, 1995.

[7] H. Bruneel. Queueing behavior of statistical multiplexers with correlated inputs. *IEEE Transactions on Communications*, COM-36(12):1339–1341, 1988.

[8] H. Bruneel. Performance of discrete-time queueing systems. *Computers and Operations Research*, 20(3):303–320, 1993.

[9] H. Bruneel and B.G. Kim. *Discrete-Time Models for Communication Systems Including ATM.* Kluwer Academic Publishers, Boston, 1993.

[10] J. A. Bucklew. *Large Deviations Techniques in Decision and Estimation.* Wiley, New York, 1990.

[11] C.-S. Chang. "Stability, queue length, and delay of deterministic and stochastic queueing networks". *IEEE Trans. Automat. Control*, 39(5):913–931, May 1994.

[12] C.-S. Chang and J. A. Thomas. "Effective bandwidth in high-speed digital networks". *IEEE JSAC*, 13(6):1091–1100, 1995.

[13] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. Weber. "Admission control and routing in ATM networks using inferences from measured buffer occupancy". *IEEE Trans. Commun.*, 43:1778–1784, 1995.

[14] C. Courcoubetis and R. Weber. "Buffer overflow asymptotics for a switch handling many traffic sources". *J. Appl. Prob.*, 33:886–903, 1996.

[15] B. Steyaert D. Fiems and H. Bruneel. Discrete-time queues with general service times and general server interruptions. *Proceedings of SPIE's International Symposium on Voice, Video and Data Communications (Boston, 6-7 November 2000).*

[16] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer, New York, 1998. First edition: Jones & Bartlett, 1993.

[17] J.H. Dshalalow. *Advances in Queueing: theory, methods and open problems.* CRC Press, Boca Raton, 1995.

[18] N. G. Duffield. "Economies of scale in queues with sources having power-law large deviation scalings". *J. Appl. Prob.*, 33:840–857, 1996.

[19] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russell, and F. Toomey. "Entropy of ATM traffic streams: A tool for estimating QoS parameters". *IEEE JSAC*, 13(6):981–990, 1995.

[20] N. G. Duffield and S. Low. "The cost of quality in networks of aggregate traffic". In *Proc. INFOCOM 1998*, pages 525–532, San Francisco, April 1998.

[21] N. G. Duffield and N. O'Connell. "Large deviations and overflow probabilities for the general single-server queue, with applications". *Math. Proc. Cambridge Philos. Soc.*, 1996.

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

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

[24] P. W. Glynn and W. Whitt. "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue". *J. Appl. Prob.*, 31A:131–156, 1994.

[25] R. Guérin, H. Ahmadi, and M. Naghshineh. "Equivalent capacity and its application to bandwidth allocation in high-speed networks". *IEEE JSAC*, 9(7):968–981, September 1991.

[26] E. Desmet H. Bruneel, B. Steyaert and G.H. Petit. Analytic derivation of tail probabilities for queue lengths and waiting times in atm multiserver queues. *European Journal of Operational Research*, 76:563–572, 1994.

[27] Q. He. Queues with marked customers. *Adv. Appl. Prob.*, 28:567–587, 1996.

[28] Q. He. Classification of Markov processes of matrix M/G/1 type with a tree structure and its applications to the MMAP[K]/G[K]/1 queue. *Stochastic Models*, 16(5):407–434, 2000.

[29] Q. He. The versatility of the MMAP[K] and the MMAP[K]/G[K]/1 queue. *Queueing Systems*, 38:397–418, 2001.

[30] Q. He and A.S. Alfa. The discrete time MMAP[K]/PH[K]/1/LCFS-GPR queue and its variants. In *Proc. of the 3rd Int. Conf. on Matrix Analytic Methods*, pages 167–190, Leuven (Belgium), 2000.

[31] Q. He and M.F. Neuts. Markov chains with marked transitions. *Stochastic Processes and their Applications*, 74:37–52, 1998.

[32] J. Hui. "Resource allocation for broadband networks". *IEEE JSAC*, 6(9):1598–1608, December 1988.

[33] J. Hui. *Switching and Traffic Theory for Integrated Broadband Networks.* Kluwer, Boston, 1990.

[34] F. P. Kelly. "Effective bandwidths at multi-class queues". *Queueing Sys.*, 9:5–15, 1991.

[35] F. P. Kelly. "Notes on effective bandwidths". In F. P. Kelly, S. Zachary, and I. Ziedens, editors, *Stochastic Networks. Theory and Applications*, volume 4 of *Royal Statistical Society Lecture Notes Series*, pages 141–168. 1996.

[36] G. Kesidis, J. Walrand, and C.-S. Chang. "Effective bandwidths for multiclass Markov fluids and other ATM sources". *IEEE/ACM Trans. Networking*, 1(4):424–428, August 1993.

[37] K. Kontovasilis and N. Mitrou. "Bursty traffic modeling and efficient analysis algorithms via fluid-flow models for ATM-IBCN". *Ann. Oper. Res.*, 49:279–323, 1994. Special Issue in Methodologies for High Speed Networks.

[38] K. Kontovasilis and N. Mitrou. "Markov modulated traffic with near complete decomposability characteristics and associated fluid queueing models". *Adv. Appl. Prob.*, 27(4):1144–1185, 1995.

[39] K. Kontovasilis and N. Mitrou. "Effective bandwidths for a class of non markovian fluid sources". *Computer Communications Review*, 27(4):263–274, 1997.

[40] L. Kosten. "Stochastic theory of data-handling systems with groups of multiple sources". In H. Rudin and W. Bux, editors, *Performance of Computer Communication Systems*, pages 321–331, Amsterdam, 1984. Elsevier.

[41] K. Kumaran and M. Mandjes. "The buffer-bandwidth trade-off curve is convex". *Queueing Sys.*, 38:471–483, 2001.

[42] D. McDonald and K. Qian. "An approximation method for complete solutions of markov-modulated fluid models". *Queueing Sys.*, 30(3–4):365–384, 1998.

[43] D. Mitra. "Stochastic theory of a fluid model of producers and consumers coupled by a buffer". *Adv. Appl. Prob.*, 20:646–676, 1988.

[44] N. Mitrou, S. Vamvakos, and K. Kontovasilis. "Modelling, parameter assessment and multiplexing analysis of bursty sources with hyperexponentially distributed bursts. *Comput. Networks ISDN Systems*, 27(7):1175–92, 1995.

[45] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach.* John Hopkins University Press, 1981.

[46] M.F. Neuts. *Structured Stochastic Matrices of M/G/1 type and their applications.* Marcel Dekker, Inc., New York and Basel, 1989.

[47] S. Wittevrongel S. De Vuyst and H. Bruneel. Statistical multiplexing of correlated variable-length packet trains : an analytic performance study. *Journal of the Operational Research Society*, 52(3):318–327, 2001.

[48] A. Simonian and J. Guibert. "Large deviations approximation for fluid queues fed by a large number of on/off sources". *IEEE JSAC*, 13(6):1017–1027, August 1995.

[49] T. E. Stern and A. I. Elwalid. "Analysis of separable Markov-modulated rate models for information-handling systems". *Adv. Appl. Prob.*, 23:105–139, 1991.

[50] B. Steyaert and H. Bruneel. Accurate approximation of the cell loss ratio in atm buffers with multiple servers. In *Performance Modelling and Evaluation of ATM Networks, Volume 1, Chapman and Hall, London, (ISBN: 0-412-71140-0)*, pages 285–296, 1995.

[51] T. Takine, B. Sengupta, and R.W. Yeung. A generalization of the matrix M/G/1 paradigm for Markov chains with a tree structure. *Stochastic Models*, 11(3):411–421, 1995.

[52] B. Van Houdt. *Performance Analysis of Contention Resolution Algorithms in Random Access Systems.* PhD thesis, University of Antwerp (UA), 2001.

[53] B. Van Houdt and C. Blondia. Stability and performance of stack algorithms for random access communication modeled as a tree structured QBD Markov chain. *Stochastic Models*, 17(3):247–270, 2001.

[54] B. Van Houdt and C. Blondia. The delay distribution of a type k customer in a first come first served MMAP[K]/PH[K]/1 queue. *Journal of Applied Probability (to appear)*, 39(1), 2002.

[55] B. Van Houdt and C. Blondia. Robustness of FS-ALOHA. In *Proc of the 4th Int. Conf. on Matrix Analytic Methods (MAM4), to appear*, Adelaide (Australia), July 2002.

[56] B. Van Houdt and C. Blondia. Robustness properties of FS-ALOHA(++): a random access algorithm for dynamic bandwidth allocation. *Journal on Special Topics in Mobile Networking and Applications (MONET) on Performance Evaluation of QoS Architectures in Mobile Networks (submitted)*, 2002.

[57] B. Van Houdt and C. Blondia. Throughput of $q$-ary splitting algorithms for contention resolution in communication networks. *To appear in Adv. in Performance Analysis*, 2002.

[58] G. de Veciana, G. Kesidis, and J. Walrand. "Resource management in wide-area ATM networks using effective bandwidths". *IEEE JSAC*, 13(6):1081–1090, 1995.

[59] B. Vinck and H. Bruneel. Relationship between delay and buffer contents in atm queues. *Electronics Letters*, 31(12):952–954, 1995.

[60] B. Vinck and H. Bruneel. Delay analysis for single server queues. *Electronics Letters*, 32(9):802–803, 1996.

[61] B. Vinck and H. Bruneel. Delay analysis of multiserver atm buffers. *Electronics Letters*, 32(5):1352–1353, 1996.

[62] A. Weiss. "A new technique for analysing large traffic systems". *Adv. Appl. Prob.*, 18:506–532, 1986.

[63] A. Weiss. "An introduction to large deviations for communication networks". *IEEE JSAC*, 13(6):938–952, August 1995.

[64] S. Wittevrongel and H. Bruneel. Deriving the tail distribution of the buffer contents in a statistical multiplexer with general heterogeneous on/off sources. In *Proceedings of the International Conference on the Performance and Management of Complex Communication Networks, PMCCN '97, (Tsukuba)*, pages 37–56, 1997.

[65] S. Wittevrongel and H. Bruneel. Correlation effects in atm queues due to data format conversions. *Performance Evaluation*, 32(1):35–56, 1998.

[66] S. Wittevrongel and H. Bruneel. Discrete-time atm queues with independent and correlated arrival streams. In *chapter 16 in : Performance Evaluation and Applications of ATM Networks, Kluwer Academic Publishers, Boston*, pages 387–412, 2000.

[67] Y. Xiong and H. Bruneel. Buffer contents and delay for statistical multiplexers with fixed-length packet-train arrival. *Performance Evaluation*, 17(1):31–42, 1993.

[68] Y. Xiong and H. Bruneel. A simple approach to obtain tight upper bounds for the asymptotic queueing behavior of statistical multiplexers with heterogeneous traffic. *Performance Evaluation*, 22(2):159–173, 1995.

[69] R.W. Yeung and A.S. Alfa. The quasi-birth-death type markov chain with a tree structure. *Stochastic Models*, 15(4):639–659, 1999.

[70] R.W. Yeung and B. Sengupta. Matrix product-form solutions for markov chains with a tree structure. *Adv. Appl. Prob.*, 26:965–987, 1994.