ELSEVIER

Performance Evaluation 23 (1995) 199–215

PERFORMANCE EVALUATION
An International Journal

Decisive arrival law parameters and a general finite capacity queueing problem

Reto Grünenfelder [a,*], Stephan Robert [b]

[a] *Alcatel STR, Friesenbergstr. 75, CH-8055 Zurich, Switzerland*

[b] *Swiss Federal Institute of Technology (EPFL), Telecommunications Laboratory, CH-1015 Lausanne, Switzerland*

Received 3 December 1992; revised 28 September 1993

Abstract

This paper addresses the problem of finding the parameters of the arrival law which most significantly influence expected occupation and loss of a finite capacity queue. The input process is supposed to be ergodic and wide sense stationary. We show that it is mostly possible to fit an MMPP(2) to the decisive parameters of observational data. Numerical examples illustrate the importance of the decisive parameters, called key parameters, and also show the accuracy of the proposed fitting procedure. Finally, in the appendix we present the solution of the finite capacity queueing problem with Special Semi Markov Process (SSMP) arrivals and a general service strategy.

Keywords: Finite capacity queue; Arrival laws; Spectral density; Experimental statistics; Fitting algorithm; Asynchronous transfer mode

1. Introduction

Multimedia communication systems move information data sets from one site to another by coordinated transmissions of data subsets. Application programs define sets of data to create multimedia information and maintain the representations associated with these sets; communication systems define collections of communication links to create multimedia streams and maintain their representations.

For the last few years the demand for multimedia services has continuously been increasing. To satisfy the provision of the Broadband Integrated Services Digital Network (B-ISDN) with high bandwidth becomes necessary. The Telecommunications Standardization Sector (TS), the former CCITT, has selected the Asynchronous Transfer Mode (ATM) as the switching technology for the B-ISDN. Today's fiber optic technology provides sufficient bandwidth for moving multimedia traffic between different locations and ATM ensures the integration and the switching of all the required different services.

* Corresponding author.

SSDI 0166-5316(94)00022-C

The multimedia traffic arriving to an ATM network is the superposition of various cell streams generated by individual input sources, such as voice, data and video. Basically, all the sources can be split into four service classes (A, B, C and D) depending on the requirements for Constant (CBR) or Variable Bit Rate (VBR) and timing relationships [17]. To provide these four service classes, four ATM Adaptation Layers (AALs) are prepared which are also based on bit rate (constant or variable) and timing relationships.

The basic parameter of the traffic descriptor of class A sources, in other words AAL1 compatible sources, is the peak cell rate. When considering, e.g. the Usage Parameter Control (UPC) at the User Network Interface (UNI) or the Quality of Service (QoS) parameter delay jitter, the maximally allowed Cell Delay Variation (CDV) has to be taken into account as well [16,18].

For all the other service classes the traffic descriptors characterize a VBR profile. Connection-oriented traffic is mostly delay sensitive and connectionless traffic is loss sensitive. To summarize, from the performance point of view delay jitter and cell loss due to UPC or congestion are the most important network performance parameters.

It is well known that traffic streams do not form a renewal process and they are mostly bursty and correlated, e.g. [14]. An accurate stochastic process characterizing such single traffic streams is the Semi Markov Process (SMP) [20] and [24]. The SMP is a Markov modulated doubly stochastic process. The Special Semi Markov Process (SSMP), a special class of the SMP, is an equivalent to the Discrete Time Markovian Arrival Process (DMAP) [4]. It is very well suited to input traffic modelling [6]. Every service, or in other words traffic stream, of a multimedia communication can accurately be fitted to an SSMP. The aggregation of all these particular services yields the traffic stream of the multimedia communication. The statistical multiplexer is a well known and accurate model for superimposing different traffic streams. Its output process, the traffic descriptor of the multimedia communication, is an SSMP as well. The superimposed input traffic stream is an SSMP and its transition matrix is given by the Kronecker product of the transition matrices of the single traffic streams. The number of arriving cells in each phase is given by the Kronecker sum [23]. Obviously, the state space of the multimedia traffic descriptor becomes enormous.

When focusing only on the most relevant network parameters, typically delay jitter and cell loss, the state space can drastically be reduced. In an ATM system the mean cell delay jitter is given by the mean waiting time in the ATM switches and multiplexers. The cell loss ratio is given by the cell loss probability. When designing or measuring ATM systems, it is important to know which traffic parameters are generic and influence most significantly the system. In Section 2 we show under very general conditions that the mean cell delay jitter and the cell loss strongly depend on the arrival law's key parameters, namely the load and the variance of each service and the spectral density at frequency zero, which is the sum over all lags of the arrival law's autocovariance function. It is evident, that load is a relevant parameter, e.g. the mean occupation in the M/D/1 queue only depends on this parameter. In Section 3 we give an algorithm for fitting observational data to an MMPP(2), while matching the key parameters. Furthermore, we show that it is not always possible to fit these three most important arrival law parameters of the aggregated traffic streams to an SSMP with only two phases while each phase describes batch arrivals. We mainly focus on an SSMP with two states since the MMPP with two states, a particular case of the SSMP, is the best known and most used traffic descriptor for

VBR sources, e.g. [1], [14] and [6]. In Section 4, we give some numerical examples for demonstrating the accuracy of the fitting method.

2. Important queueing parameters

In this section we consider the mean cell delay and the cell loss during a busy period of a queueing system. A busy period is defined to begin with the arrival of a cell to an idle queueing system (time $-\tau$) and to end when it next becomes idle (time τ). Let $N(t)$ denote the number of cells in the queueing system at time t, where $-\tau < t < \tau$, $N(t) > 0$ and $N(-\tau) = N(\tau) = 0$ and $U(t)$ the number of arriving minus the number of served cells at time t. Furthermore, we assume the queueing system's capacity to be limited to c places and that there are k excess periods during the busy period $-\tau < t < \tau$, $t_0 = -\tau$ and $t_{2k+1} = \tau$. The contents $N(t)$ is then given by

$$N(t) = \begin{cases} \int_{t_0}^{t} U(s)\,\mathrm{d}s, & -\tau = t_0 \le t < t_1, \\ c, & t_{2j-1} \le t < t_{2j},\ 1 \le j \le k, \\ c + \int_{t_{2j}}^{t} U(s)\,\mathrm{d}s, & t_{2j} \le t < t_{2j+1},\ 1 \le j \le k. \end{cases} \tag{1}$$

For the definition of the times t_i, $0 \le i \le 2k+1$ we refer to Fig. 1. Let $M(t)$ denote the cell loss at time t, it is given by

$$M(t) = \begin{cases} 0, & t_{2j} \le t < t_{2j+1},\ 0 \le j \le k, \\ \int_{t_{2j-1}}^{t} U(s)\,\mathrm{d}s, & t_{2j-1} \le t < t_{2j},\ 1 \le j \le k. \end{cases} \tag{2}$$

We now focus only on the queueing system occupation $N(t)$. Later on, it will be shown that similar arguments can also be applied to the loss $M(t)$. Define the function $\Psi(\omega)$ by

$$\Psi(\omega) = \frac{1}{2\tau} \int_{-\tau}^{\tau} N(t)\, \mathrm{e}^{-\mathrm{i}\omega t}\, \mathrm{d}t \tag{3}$$

where $\Psi(0) = E[N \mid -\tau < t < \tau]$. E is the operator of the expected value. An excess period is defined to begin when the queueing system starts to be completely occupied, e.g. t_1 in Fig. 1,

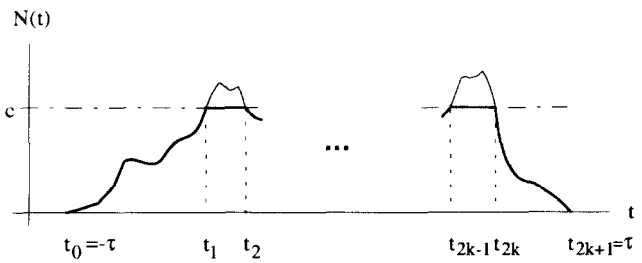


Fig. 1. Queueing system occupation versus time.

and to end when it next becomes only partially filled, e.g. t_2 in Fig. 1, this means $N(t_1-)<c$ and $N(t_2+)<c$.

After substituting (1) into (3) and partial integration we obtain

$$\Psi(\omega)=\frac{-1}{\mathrm{i}2\omega\tau}\int_{-\tau}^{\tau}U(t)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}t=\frac{-1}{\mathrm{i}2\omega\tau}\sum_{j=0}^{k}\int_{t_{2j}}^{t_{2j+1}}U(t)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}t. \tag{4}$$

It is useful to remember that

$$\mathrm{d}N(t)=\begin{cases}0, & t_{2j-1}\le t<t_{2j},\ 1\le j\le k,\\ U(t)\,\mathrm{d}t, & t_{2j}\le t<t_{2j+1},\ 0\le j\le k.\end{cases}$$

We next introduce the characteristic function $\chi_{[a,b]}$, which is 1 for $a\le t\le b$ and 0 otherwise. Let the Fourier transform $\mathcal{F}$ of a function $f(t)$ be defined as usual in signal processing, namely by

$$\mathcal{F}\{f(t)\}=\int_{-\infty}^{\infty}f(t)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}t.$$

Hence (4) can be expressed as Fourier transform, this yields

$$\Psi(\omega)=\frac{-1}{\mathrm{i}2\omega\tau}\sum_{j=0}^{k}\int_{-\tau}^{\tau}\chi_{[t_{2j},t_{2j+1}]}U(t)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}t=\frac{-1}{\mathrm{i}2\omega\tau}\sum_{j=0}^{k}\mathcal{F}\left\{\chi_{[t_{2j},t_{2j+1}]}U(t)\right\}. \tag{5}$$

We now calculate $\Psi(\omega)\overline{\Psi}(\omega)$, which is for $\omega=0$ the square of the expected value of the queueing system occupation N:

$$\begin{aligned}\Psi(\omega)\overline{\Psi}(\omega)&=\frac{1}{4\tau^2\omega^2}\sum_{j,l=0}^{k}\int_{-\tau}^{\tau}\mathrm{d}t\int_{-\tau}^{\tau}\mathrm{d}\tilde{t}\,\chi_{[t_{2j},t_{2j+1}]}\chi_{[\tilde{t}_{2l},\tilde{t}_{2l+1}]}U(t)U(\tilde{t})\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{e}^{\mathrm{i}\omega\tilde{t}}\\&=\frac{1}{4\tau^2\omega^2}\sum_{j=0}^{k}\int_{-\tau}^{\tau}\mathrm{d}t\int_{-\tau}^{\tau}\mathrm{d}\tilde{t}\,\chi_{[t_{2j},t_{2j+1}]}U(t)U(\tilde{t})\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{e}^{\mathrm{i}\omega\tilde{t}}.\end{aligned} \tag{6}$$

Changing the variables to $v=t-\tilde{t}$ and $t=t$ yields

$$\Psi(\omega)\overline{\Psi}(\omega)=\frac{1}{4\tau^2\omega^2}\sum_{j=0}^{k}\int_{-2\tau}^{2\tau}\mathrm{e}^{-\mathrm{i}\omega v}\,\mathrm{d}v\int_{-\tau}^{\tau}\chi_{[t_{2j},t_{2j+1}]}U(t)U(t-v)\,\mathrm{d}t. \tag{7}$$

In the following we assume the process $U(t)$ to be ergodic, this means

$$\frac{1}{2\tau}\int_{-\tau}^{\tau}\sum_{j=0}^{k}\chi_{[t_{2j},t_{2j+1}]}U(t)U(t-v)\,\mathrm{d}t=E\left[U(t)U(t-v)\,|\,t\in\Omega\right] \tag{8}$$

where

$$\Omega=\bigcup_{j=0}^{k}\left\{t\colon t_{2j}\le t\le t_{2j+1}\right\}.$$

Furthermore, we assume the process $U(t)$ to be wide sense stationary, which means that $E[U(t)]=\eta$ is independent of t, and that the autocorrelation $E[U(t+v)U(t)]=R_U(v)$ only depends on the lag v. Obviously, $R_U(v)$ is symmetric, that means $R_U(v)=R_U(-v)$. Hence, expression (8) becomes $R_U(v)$.

The spectral density $\Phi_U(\omega)$ of the process $U(t)$ is the Fourier transform of the autocovariance function $C_U(v) = R_U(v) - \eta^2$ [10]. With this and the convolution theorem for Fourier transforms, expression (7) becomes

$$\Psi(\omega)\overline{\Psi}(\omega) = \frac{1}{2\tau\omega^2}\mathcal{F}\{\chi_{[-2\tau,2\tau]}\} * \left(\Phi_U(\omega) + \mathcal{F}\{\eta^2\}\right)$$
$$= \frac{1}{\tau\omega^2}\int_{-\infty}^{\infty} \frac{\sin(2\alpha\tau)}{\alpha}\left(\Phi_U(\omega-\alpha) + 2\pi\eta^2\delta(\omega-\alpha)\right) \mathrm{d}\alpha. \tag{9}$$

In the following we approximately evaluate the integral of (9) at $\omega = 0$. The term $\sin(2\alpha\tau)/\alpha$ has its main lobe at $\alpha = 0$. The greater τ the more important becomes this main lobe. Its width decreases with increasing τ. Or in other words, the greater τ the more important become low frequencies. We therefore assume that ω is proportional to $1/\tau$. This yields for small α that

$$1/\tau\omega^2 \,\tilde{\alpha}\, \omega/(\omega^2 + \alpha^2) \tag{10}$$

where $\tilde{\alpha}$ means approximately proportional to.

Equation (9) can be given as an approximation, when substituting (10) into (9), taking $\omega \to 0$, and limit and integration can be interchanged to obtain

$$\Psi(0)\overline{\Psi}(0) \,\tilde{\alpha} \int_{-\infty}^{\infty} \lim_{\omega\to 0}\left(\frac{\omega}{\omega^2+\alpha^2}\frac{\sin(2\alpha\tau)}{\alpha}\left(\Phi_U(\omega-\alpha) + 2\pi\eta^2\delta(\omega-\alpha)\right)\right) \mathrm{d}\alpha. \tag{11}$$

Since

$$\lim_{\omega\to 0}\frac{\omega}{\omega^2+\alpha^2} = \pi\delta(\alpha) \quad \text{and} \quad \lim_{x\to 0}\frac{\sin(2\tau x)}{x} = 2\tau$$

we obtain

$$\Psi(0)\overline{\Psi}(0) = \left(E[N \mid -\tau < t < \tau]\right)^2 \,\tilde{\alpha}\, 2\pi\tau\left(\Phi_U(0) + 2\pi\eta^2\right). \tag{12}$$

Equation (12) shows that during a busy period the squared expected value of the queueing system occupation is approximately proportional to an expression in which the following parameters are involved:

- the expected value of the process $U(t)$, that means the difference between arrival rate and service rate,
- the spectral density Φ_U at frequency 0 of the process $U(t)$,
- the duration of the busy period, given as 2τ.

The spectral density at frequency zero is equal to the integral over all lags of the autocovariance function of the process $U(t)$. This shows that the occupation of the queueing system does not depend on the detailed but the *global behaviour of the autocovariance function* of the process $U(t)$. For ATM systems typically, the service rate is constant and hence, only the spectral density at frequency zero of the *arrival process* influences the queueing system occupation. Furthermore, expression (12) shows that the expected value of $U(t)$, the difference of mean arrival and service arrival rate, also decisively influences the queueing system occupation. The $\Phi_U(0)$ becomes for non-correlated arrivals equal to the variance. That means $\Phi_U(0)$ consists of the variance plus an extra term which characterizes the correlation. Thus, the influence of correlation on the queueing behaviour becomes more evident. It is therefore

suitable to consider also the variance of U, which gives an integral information about the whole spectral density, as a key parameter. The duration of the busy period strongly depends on the fluctuations of the service process of the queueing system [11]. Therefore, it is obvious to replace the relevant parameter busy period by the relevant parameter, variance of the process $U(t)$. Hence, the key parameters for the mean occupation — also for the loss as seen later on — of a queueing system are

- the expected arrival and the expected service rate, more precisely the expectation of the process $U(t)$,
- the variance of the process $U(t)$,
- the spectral density at frequency zero of the process $U(t)$.

A result which is similar to that of the mean occupation can now be derived easily for the mean cell loss during the busy period. Let $\phi(\omega)$ denote the Fourier transform of the mean cell loss. The busy period of the loss starts, see Fig. 1, at the time t_1 and ends up with t_{2k}. In the following, these instants will be named as $-\tau$ and τ respectively. The Fourier transform of the mean loss is then given by

$$\phi(\omega) = \frac{-1}{\mathrm{i}2\omega\tau} \sum_{j=1}^{k} \int_{t_{2j-1}}^{t_{2j}} U(t)\, \mathrm{e}^{-\mathrm{i}\omega\tau}\, \mathrm{d}t. \tag{13}$$

This equation is form-invariant to (4). Finally, we obtain for the squared mean loss a similar expression as for the mean occupation.

$$\phi(0)\overline{\phi}(0) = \left(E[M \mid -\tau < t < \tau]\right)^2 \tilde{\alpha}\, 2\pi\tau\left(\Phi_U(0) + 2\pi\eta^2\right). \tag{14}$$

It is important to point out that the factor 2τ in (14) corresponds to the elapsed time between the first and the last loss during a busy period. Furthermore, it is very important to note that it is an approximate proportionality. In other words, both equations (12) and (14) do not serve for the computation of the mean occupation and the mean cell loss in a queueing system, but they clearly show the importance of parameters. These key parameters are, as already pointed out above, the spectral density at frequency zero of the process $U(t)$, the variance, the mean arrival rate and the mean service rate. This is valid for the mean occupation as well as the mean loss. The analytical study of the mean occupation of the Σ2SM/D/1 queue [21] clearly proves that the mean occupation only depends on the key parameters mentioned above. (2SM means two state Markov process, which is a generalization of the Bernoulli process.)

3. The fitting procedure

The probably best known process characterizing non-renewal arrivals is the Markov Modulated Poisson Process (MMPP) with two phases. Several fitting procedures for the MMPP(2) are given in literature, e.g. [1,6,14,22,25]

The procedure in [22] focuses on the parameter estimation of an MMPP(2) based on observational interarrival times. The method bases on iteration and is motivated by the maximum likelihood estimation. The method in [6] allows observational data of a process to be completely fitted to the distribution function of an SSMP(2). It is furthermore possible to fit the

first lag of the autocovariance function. This method assumes that the arrivals in state i occur according to a general process (not Poisson). The methods of [14] and [25] are similar. Both fit the mean arrival rate and the long term variance-to-mean ratio of the number of arrivals in $(0, t)$ with $t \to \infty$. Furthermore, [14] proposes to fit the variance to mean ratio and the third moment, both of the number of arrivals during a finite interval. [25] proposes to fit the covariance of the number of arrivals of two consecutive infinitely long time intervals and the squared coefficient of variation of the arrival times. The common goal of the four last procedures is that they propose to accurately fit the arrival process of a queueing system.

The main goal in the fitting procedure in [1] is to fit the parameters of the MMPP(2), such that important effects of the queueing behaviour can be seen, e.g. cell loss versus buffer capacity, cell and burst scale behaviour. The procedure consists of splitting the superimposed traffic stream into an underload and an overload phase. During underload, the instantaneous arrival rate is smaller than the mean arrival rate. The definition of overload is analogous. It has been shown, [1, Fig. 1], that such a fitting procedure is more accurate than the former ones.

In the following, we present a technique for approximating superimposed traffic streams to an SSMP(2). We assume that when the SSMP is in state $Y_t, i \in \{1, 2\}$, arrivals occur according to a Poisson process of rate λ_i. The modulator's transition probabilities are defined as

$$p = \Pr(Y_{t+1} = 2 \mid Y_t = 1) \quad \text{and} \quad q = \Pr(Y_{t+1} = 1 \mid Y_t = 2).$$

The aim of our fitting procedure is to fit the most relevant parameters, that means those which most decisively influence the performance of the queueing system. It can be shown [6] that such an SSMP(2) is equivalent to the MMPP(2).

Let X denote the random variable characterizing the number of arrivals during one time slot. A time slot is a fixed length interval. Its length is equal to the time it takes to transmit one cell. For the considered SSMP(2) the following set of equations can easily be derived:

$$E[X] = \frac{1}{1+\alpha}(\lambda_1 + \alpha\lambda_2), \quad \alpha = p/q, \tag{15}$$

$$\mathrm{Var}[X] = E[X] + \frac{\alpha}{(1+\alpha)^2}(\lambda_1 - \lambda_2)^2, \quad \alpha = p/q, \tag{16}$$

$$E[X^3] = -2E[X] + 3E[X^2] + \frac{1}{1+\alpha}(\lambda_1^3 + \alpha\lambda_2^3), \quad \alpha = p/q, \tag{17}$$

$$\Phi_X(0) = E[X] + \left(\frac{2}{p+q} - 1\right)(\mathrm{Var}[X] - E[X]). \tag{18}$$

Now we consider the statistical problem of fitting an SSMP(2) with Poisson arrivals, in the following called MMPP(2), to observational data. Let $\mu_X^{(n)}$ denote the nth moment of the observational data, σ_X^2 the sample variance and $\hat{\Phi}_X(0)$ the spectral density at frequency zero. Techniques for measuring on-line such data are presented in [12]. The knowledge of the key parameters is of particular interest for testing performance and network behaviour of the B-ISDN.

From Eqs. (16) and (18) a priori follows that observational data can only be matched accurately to an MMPP(2) if $\sigma_X^2 \geq \mu_X^{(1)}$ and $\hat{\Phi}_X(0) \geq \mu_X^{(1)}$ respectively. The first restriction obviously comes from the assumption that arrivals occur for each state according to a Poisson

process. The second limitation is valid for all SSMPs with a two state modulator. Equation (18) is independent of the Poisson assumption, since the spectral density of any SSMP only depends on the modulator and the mean rate of each state. To summarize, relevant statistical parameters of observational data cannot always be matched to an MMPP(2) or even the more general SSMP(2). To demonstrate the relevance of the key parameters, found in the last section, we stick to the MMPP(2). The following proposition is useful when matching observational data to an MMPP(2). It gives a fundamental relationship between the transition probabilities p and q.

Proposition. *All the pairs of transition probabilities* $(p, q) \in \mathscr{A}$ *match the observed mean and the observed variance to an* MMPP(2) *where*

$$\mathscr{A} = \left\{ (p, q): 1 \geq p \geq q\delta \geq 0,\ \delta = \frac{\sigma_X^2 - \mu_X^{(1)}}{\left(\mu_X^{(1)}\right)^2} \right\} \tag{19}$$

and

$$0 \leq q \leq 1 \ \textit{iff}\ 0 \leq \delta \leq 1, \qquad 0 \leq q < 1/\delta \ \textit{iff}\ \delta > 1.$$

Proof. Define

$$M \equiv \sqrt{\frac{(p+q)^2}{pq}(\mathrm{Var}[X] - E[X])},$$

which is, due to Eq. (16), non-negative. Without loss of generality we assume that $\lambda_1 \geq \lambda_2$. We then obtain from (16)

$$\lambda_1 = \lambda_2 + M. \tag{20}$$

Finally, (15) yields

$$\lambda_2 = E[X] - \frac{q}{p+q}M \geq 0 \tag{21}$$

from which we obtain (19). □

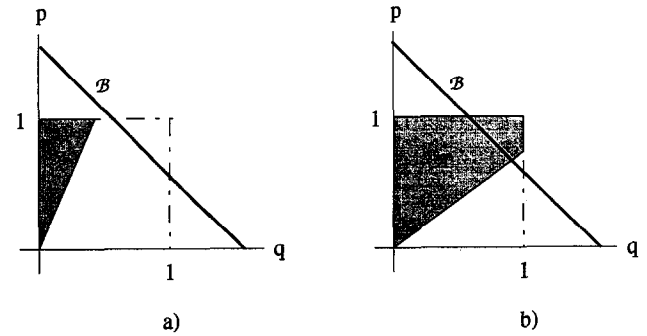


Fig. 2. Comparison of sets $\mathscr{A}$ and $\mathscr{B}$. For case (a) an empty intersection of $\mathscr{A}$ and $\mathscr{B}$ is possible. In case (b) the intersection never becomes empty.

A further restriction for the choice of the pairs of transition probabilities is given by Eq. (18). Given that the observed mean and variance match with the MMPP(2), the observed spectral density at frequency zero can be matched as long as $(p, q) \in \mathscr{B}$ where

$$\mathscr{B} = \left\{ (p, q):\ p + q = \frac{2(\sigma_X^2 - \mu_X^{(1)})}{\hat{\Phi}_X(0) + \sigma_X^2 - 2\mu_X^{(1)}},\ 0 \le p, q \le 1 \right\}. \tag{22}$$

Obviously, the set $\mathscr{A} \cap \mathscr{B}$ may be empty (see Fig. 2), which means that either mean and variance or the spectral density at frequency zero do not matched with the MMPP(2). In the case of $\mathscr{A} \cap \mathscr{B} \neq \emptyset$, the optimal (p, q) can be found by e.g. minimizing

$$|(E[X^3]) - \mu_X^{(3)}|^2.$$

To summarize, the algorithm for matching the key parameters of observational data to an MMPP(2) is given in the list below.

Algorithm

1. Determine the sets $\mathscr{A}$ given by (19) and $\mathscr{B}$ given by (22).
2. If $\mathscr{A} \cap \mathscr{B} \neq \emptyset$ goto 3, otherwise goto 4.
3. Compute the optimal pair (p, q) based on (17), goto 5 (for the rates use the expressions (20) and (21)).
4. Compute the optimal pair (p, q) such that the matching error between observed and theoretical spectral density becomes minimal. Remark: For an approximately minimal error choose the pair $(p, q) \in \mathscr{A}$ which has the smallest distance to the set $\mathscr{B}$.
5. Compute λ_2 given by (21) and λ_1 given by (20).

Table 1
Parameter choice for case study I

Original	$A = \begin{pmatrix} 0.1 & 0.8 & 0.1 \\ 0.2 & 0.1 & 0.7 \\ 0.8 & 0.1 & 0.1 \end{pmatrix}$	$\lambda_1 = 0.1$ $\lambda_2 = 2.0$ $\lambda_3 = 0.25$	$\mu_X^{(1)} = 0.8$ $\sigma_X^2 = 1.56$ $\mu_X^{(3)} = 7.88$ $\hat{\Phi}_X(0) = 0.98735$
Fit to MMPP(2)	$A = \begin{pmatrix} 0 & 1 \\ 0.604 & 0.396 \end{pmatrix}$	$\lambda_1 = 1.92173$ $\lambda_2 = 0.122475$	$E[X] = 0.8$ $\mathrm{Var}[X] = 1.56$ $E[X^3] = 7.673$ $\hat{\Phi}_X(0) = 0.98735$
Poisson	$A = 1$	$\lambda_1 = 0.8$	$E[X] = 0.8$ $\mathrm{Var}[X] = 0.8$ $E[X^3] = 3.232$ $\hat{\Phi}_X(0) = 0.8$

Table 2
Parameter choice for case study II

Original	$A = \begin{pmatrix} 0.1 & 0.8 & 0.1 \\ 0.2 & 0.1 & 0.7 \\ 0.8 & 0.1 & 0.1 \end{pmatrix}$	$\lambda_1 = 0.0001$ $\lambda_2 = 0.85$ $\lambda_3 = 0.025$	$\mu_X^{(1)} = 0.3$ $\sigma_X^2 = 0.459$ $\mu_X^{(3)} = 1.277$ $\hat{\Phi}_X(0) = 0.340$
Fit to MMPP(2)	$A = \begin{pmatrix} 0 & 1 \\ 0.568 & 0.432 \end{pmatrix}$	$\lambda_1 = 0.833$ $\lambda_2 = 0$	$E[X] = 0.3$ $\mathrm{Var}[X] = 0.549$ $E[X^3] = 1.275$ $\hat{\Phi}_X(0) = 0.3438$
Poisson	$A = 1$	$\lambda_1 = 0.3$	$E[X] = 0.3$ $\mathrm{Var}[X] = 0.3$ $E[X^3] = 0.597$ $\hat{\Phi}_X(0) = 0.3$

4. Numerical examples

In Tables 1–3 and Figs. 3–5 numerical examples are given that demonstrate the relevance of the key parameters found in Section 2. In the examples we compute with the MBH algorithm the expected loss of the SSMP(2)/D/1/c. The appendix gives an overview on how to compute efficiently the SSMP(n)/G/1/c queueing system. We assume that the observational data will

Table 3
Parameter choice for case study III

Original	$A = \begin{pmatrix} 0.05 & 0.95 & 0 \\ 1E-4 & 0.99989 & 1E-5 \\ 0 & 0.95 & 0.05 \end{pmatrix}$	$\lambda_1 = 50.0$ $\lambda_2 = 0.50$ $\lambda_3 = 5.0$	$\mu_X^{(1)} = 0.505$ $\sigma_X^2 = 0.763$ $\mu_X^{(3)} = 15.328$ $\hat{\Phi}_X(0) = 0.790$
Fit to MMPP(2)	$A = \begin{pmatrix} 1-9.1E-5 & 9.1E-5 \\ 0.950017 & 0.049983 \end{pmatrix}$	$\lambda_1 = 0.500283$ $\lambda_2 = 52.386$	$E[X] = 0.505$ $\mathrm{Var}[X] = 0.763$ $E[X^3] = 15.953$ $\hat{\Phi}_X(0) = 0.790$
Poisson	$A = 1$	$\lambda_1 = 0.3$	$E[X] = 0.505$ $\mathrm{Var}[X] = 0.505$ $E[X^3] = 1.399$ $\hat{\Phi}_X(0) = 0.505$

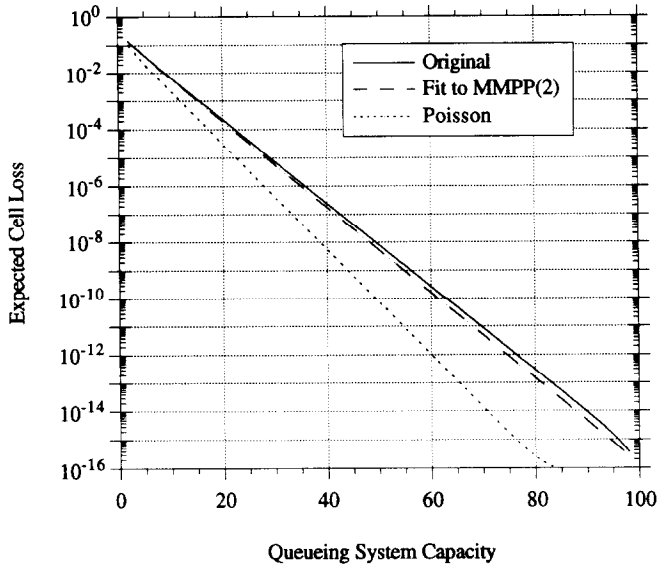


Fig. 3. Expected cell loss versus queueing system capacity for case study I.

be produced by an SSMP(3) with Poisson arrivals. We fit it with an MMPP(2) while matching the key parameters with the presented fitting algorithm.

The comparison of the expected loss for the observational data (SSMP(3)) with the matched MMPP(2) shows that the proposed fitting guarantees good results. Hence, the relevance of the key parameters, found in Section 2, has been validated.

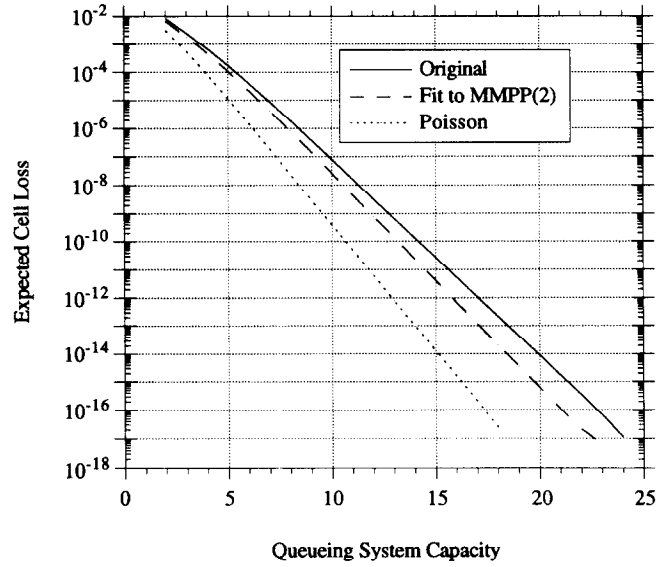


Fig. 4. Expected cell loss versus queueing system capacity for case study II.

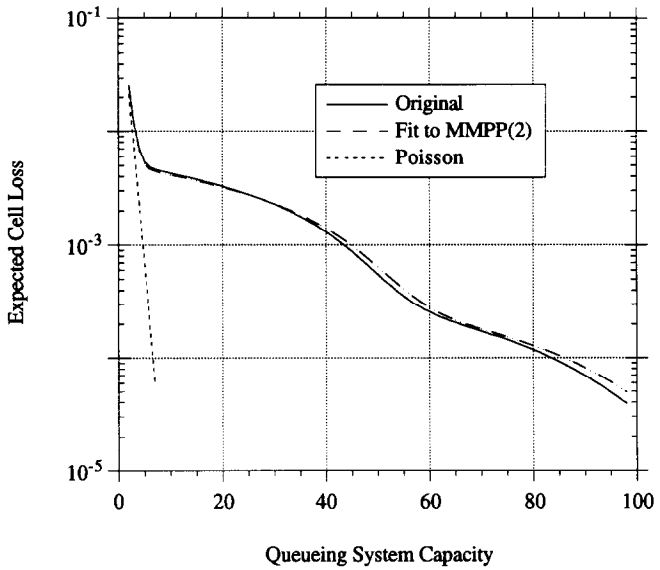


Fig. 5. Expected cell loss versus queueing system capacity for case study III.

5. Conclusions

Firstly, we have considered the expected occupation and the expected loss of a finite queue during a busy period. This consideration has shown that the queueing behaviour strongly depends on three parameters, the so-called key parameters, which are the expected arrival rate, the expected service rate and the spectral density at the frequency zero of the difference between arrivals and service. This difference is supposed to be ergodic and wide sense stationary. A typical example of such a process is the SSMP. The motivation for using a SSMP for describing traffic has been given by a multimedia. Furthermore, the aggregation of ATM traffic can be described easily by such a process. The disadvantage is that the state space of the underlying Markov chain dramatically grows. To exploit the results about the key parameters of Section 2, we presented in Section 3 a method to fit the key parameters of observational data to a MMPP(2). The importance of the key parameters has been illustrated by some numerical examples, which make evident that only the key parameters have a relevant influence on the queueing behaviour. The use of the fitting algorithm is manifold, e.g., SSMPs with a tremendous state space can be reduced significantly, observational data can be fitted accurately. The computation time of the queueing problem with MMPP(2) batch arrivals is very short. The findings can serve as a framework for an engineering tool which allows approximate calculation of loss and delay jitter in an ATM network. The complete algorithm for solving the SSMP/G/1/c is given in the appendix. Furthermore, the knowledge of the key parameters brings some basic and important insight for the parameter choice of the performance test and the end-to-end test of B-ISDN.

Appendix A

In this appendix we outline a general algorithm for solving the SSMP/G/1 finite capacity queue. A similar queueing problem is presented in [2], where the arrival law is a Markovian Arrival Process (MAP). The waiting time distribution of the SSMP/G/1 queue with exponential interarrivals in each state is presented in [9]. The finite capacity queueing problem with SMP arrivals and deterministic server can be found in [15]. In the context of this contribution we focus on buffer occupation and loss probability. The calculation of the waiting time distribution is straightforward. The presented method was pioneered by [5] and [13] and has been used successfully to solve a large number of stochastic queueing models. The results presented here are not new, the purpose of the presentation is to make this powerful tool available for a wider audience.

The arrival process is supposed to be a Special Semi-Markov Process (SSMP) with batch arrivals. It is well suited for characterizing the arrivals to a queue of a statistical multiplexer. Our SSMP is similar to the Discrete-Time Batch Markovian Arrival Process (DBMAP) [3]. Due to the nature of an SSMP we suppose the time to be slotted, with the slot length equal to the time it takes to transmit one cell. Time slots are indexed by t. The arrivals of cells are modulated by an n-state discrete time Markov chain (modulator) with transition probabilities $a_{ij} = \Pr(Y_{t+1} = j \mid Y_t = i)$. In the modulator's state i, $i \in \{1, \ldots, n\}$ cells are generated due to a general process β_i. Let X_t denote the number of cells that arrive during the interval $[t-l, t)$. Let ϕ_{ij} denote the probability of having j cells given that the modulator's state is i, more specifically $\phi_{ij} = \Pr(X_t = j \mid Y_t = i)$. The interval $[t-l, t)$ is to be referred to as the tth slot.

The queue itself is a single priority FCFS (First Come First Served) queue with a single server. The queueing system (queue plus server) consists of c places. The arrival of cells are assumed at the beginning of a slot. Any departure from the queuing system is assumed to take place at the end of a slot.

Let N_t denote the number of cells in the queueing system and R_t the residual service time both at time t. The service times H are positive integer multiples of time slots and $h_r = \Pr(H = r)$. If $N_t = 0$, R_t is defined to be zero. Note that the triple-variate process $\{N_t, R_t, Y_t\}$ with state space

$$\{\{(m, r, j) \mid 1 \le m \le c,\ 1 \le r < \infty,\ j \in \{1, \ldots, n\}\} \cup \{(0, 0, j) \mid j \in \{1, \ldots, n\}\}\}$$

forms a Markov chain.

To obtain the difference equation of the occupation N_t we follow [5] and [13] relating the probabilities at $t+1$ to those at t. This leads to

$m \ge 1$:

$$\begin{aligned}
&\Pr(N_{t+1} = \min\{m, c\} \cap R_{t+1} = r \cap Y_{t+1} = j \mid N_0 = 0 \cap Y_0 = i) \\
&\quad = \sum_{s=1}^{n} \Bigg\{ \sum_{k=1}^{\min\{m,c\}} \Pr\big((N_t = k \cap R_t = r+1 \cap Y_t = s \mid N_0 = 0 \cap Y_0 = i) \\
&\qquad\qquad \cap (X_t = m - k \mid Y_t = s) \cap (Y_{t+1} = j \mid Y_t = s) \big) \\
&\qquad + \sum_{k=1}^{\min\{m+1,c\}} \Pr\big((N_t = k \cap R_t = 1 \cap Y_t = s \mid N_0 = 0 \cap Y_0 = i) \\
&\qquad\qquad \cap (X_t = m - k + 1 \mid Y_t = s) \cap (Y_{t+1} = j \mid Y_t = s) \cap (H = r) \big)
\end{aligned}$$

$$+\Pr\big((N_t=0\cap Y_t=s\mid N_0=0\cap Y_0=i)$$

$$\cap(X_t=m\mid Y_t=s)\cap(Y_{t+1}=j\mid Y_t=s)\cap(H=r)\big)\Big\};$$

$m=0$:

$$\Pr(N_{t+1}=0\cap Y_{t+1}=j\mid N_0=0\cap Y_0=i)$$

$$=\sum_{s=1}^{n}\{\Pr\big((N_t=1\cap R_t=1\cap Y_t=s\mid N_0=0\cap Y_0=i)\cap(X_t=0\mid Y_t=s)$$

$$\cap(Y_{t+1}=j\mid Y_t=s)\big)$$

$$+\Pr\big((N_t=0\cap Y_t=s\mid N_0=0\cap Y_0=i)\cap(X_t=0\mid Y_t=s)\cap(Y_{t+1}=j\mid Y_t=s)\big)\}. \tag{A1}$$

When introducing the following notations:

$$p_{ij}(\min\{m,c\},r,t+1)\equiv\Pr(N_t=\min\{m,c\}\cap R_t=r\cap Y_t=j\mid N_0=0\cap Y_0=i),$$

$$p_{ij}(0,t+1)\equiv\Pr(N_t=0\cap Y_t=j\mid N_0=0\cap Y_0=i), \tag{A2}$$

the difference equation (A1) can be rewritten as follows:

$m\geq 1$:

$$p_{ij}(\min\{m,c\},r,t+1)=\sum_{s=1}^{n}\Bigg(\sum_{k=1}^{\min\{m,c\}}p_{is}(k,r+1,t)\phi_{s,m-k}a_{sj}$$

$$+h_r\sum_{k=1}^{\min\{m+1,c\}}p_{is}(k,1,t)\phi_{s,m-k+1}a_{sj}+h_r p_{is}(0,t)\phi_{sm}a_{sj}\Bigg);$$

$m=0$:

$$p_{ij}(0,t+1)=\sum_{s=1}^{n}(p_{is}(1,1,t)+p_{is}(0,t))\phi_{s,0}a_{sj}. \tag{A3}$$

We assume the existence of the steady state probabilities for $t\to\infty$, that means that the Markov chain is homogeneous and irreducible. Then we obtain

$m\geq 1$:

$$\lim_{t\to\infty}\big(p_{ij}(\min\{m,c\},r,t+1)\big)$$

$$=p_j(\min\{m,c\},r)=\sum_{s=1}^{n}\Bigg(\sum_{k=1}^{\min\{m,c\}}p_s(k,r+1)\phi_{s,m-k}a_{sj}$$

$$+h_r\sum_{k=1}^{\min\{m+1,c\}}p_s(k,1)\phi_{s,m-k+1}a_{sj}+h_r p_s(0)\phi_{sm}a_{sj}\Bigg);$$

$m=0$:

$$\lim_{t\to\infty}\big(p_{ij}(0,t+1)\big)=p_j(0)=\sum_{s=1}^{n}(p_s(1,1)+p_s(0))\phi_{s,0}a_{sj}. \tag{A4}$$

Taking $t \to \infty$ yields that (A4) becomes independent of the initial state of the Markov chain. In the following we assume the residual service time to be independent of the system occupation, that means for $r \geq 1$: $\boldsymbol{p}_j(k, r) = \boldsymbol{p}_j(k)\gamma_r$ with $\gamma_r = \Pr(R = r)$.
Remember that $r \geq 0$ if at least one cell is in the queueing system and that $\sum_{r=1}^{\infty}\gamma_r = 1$. (A4) then becomes

$m \geq 1$:

$$\boldsymbol{p}_j(\min\{m, c\}) = \sum_{s=1}^{n}\left(\sum_{k=1}^{\min\{m, c\}} \boldsymbol{p}_s(k)\phi_{s,m-k}a_{sj}(1-\gamma_1) + \sum_{k=1}^{\min\{m+1,c\}} \boldsymbol{p}_s(k)\phi_{s,m-k+1}a_{sj}\gamma_1 + \boldsymbol{p}_s(0)\phi_{sm}a_{sj}\right);$$

$m = 0$:

$$\boldsymbol{p}_j(0) = \sum_{s=1}^{n}(\gamma_1\boldsymbol{p}_s(1) + \boldsymbol{p}_s(0))\phi_{s,0}a_{sj}. \tag{A5}$$

Let

$$\boldsymbol{X} = (\boldsymbol{p}_1(0), \ldots, \boldsymbol{p}_n(0), \boldsymbol{p}_1(1), \ldots, \boldsymbol{p}_n(1), \ldots, \boldsymbol{p}_1(\min\{m, c\}), \ldots, \boldsymbol{p}_n(\min\{m, c\}))$$

denote the vector of the steady state probabilities and

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \boldsymbol{B}_2 & \cdots & \boldsymbol{B}_{c-1} & \sum_{l\geq c}\boldsymbol{B}_l \\ \boldsymbol{H}_0 & \boldsymbol{H}_1 & \boldsymbol{H}_2 & \cdots & \boldsymbol{H}_{c-1} & \sum_{l\geq c}\boldsymbol{H}_l \\ 0 & \boldsymbol{H}_0 & \boldsymbol{H}_1 & \cdots & \boldsymbol{H}_{c-2} & \sum_{l\geq c-1}\boldsymbol{H}_l \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & \boldsymbol{H}_0 & \sum_{l\geq 1}\boldsymbol{H}_l \end{pmatrix} \tag{A6}$$

the transition matrix of the queueing system. With (A6), the set of equations (A5) can be written as $\boldsymbol{X} = \boldsymbol{X}\cdot\boldsymbol{Q}$. The blocks in the matrix $\boldsymbol{Q}$ are given as follows:

$$(\boldsymbol{B}_k)_{ij} = a_{ij}\phi_{ij}, \quad 0 \leq k \leq c-1 \text{ and } 1 \leq i, j \leq n,$$

$$\left(\sum_{l\geq c}\boldsymbol{B}_l\right)_{ij} = a_{ij} - \sum_{k=0}^{c-1}(\boldsymbol{B}_k)_{ij}, \quad 1 \leq i, j \leq n$$

$$(\boldsymbol{H}_k)_{ij} = a_{ij}(\gamma_1\phi_{ik} + (1-\gamma_1)\phi_{i,k-1}), \quad 0 \leq k \leq c-1, 1 \leq i, j \leq c \text{ and } \phi_{i,-1} \equiv 0,$$

$$\left(\sum_{l\geq c}\boldsymbol{H}_l\right)_{ij} = a_{ij} - \sum_{k=0}^{c-1}(\boldsymbol{H}_k)_{ij}, \quad 1 \leq i, j \leq n. \tag{A7}$$

Hence we have shown that the buffer state can be represented by a finite state Markov chain whose transition matrix is upper block-Hessenberg, namely, with no blocks below the one

parallel to the diagonal. For such matrices, the steady state probability vector X can be efficiently computed by the Markov Block-Hessenberg (MBH) algorithm [8] and [19]. Furthermore, MBH delivers (based on X) the probability of the cell loss due to buffer overflow.

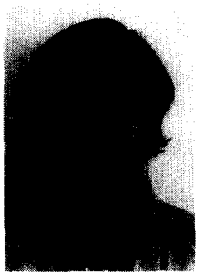
The transition matrix Q consists of only $3c + 2$ different blocks, where each is a substochastic matrix of dimension $n \times n$ (n is the number of states of the modulator). The key feature of the MBH algorithm is that it deals only with substochastic matrices. Therefore, it is numerically more stable than classical iteration methods such as Gauss–Seidel and overrelaxation. The computation time of the SSMP(2)/D/1/100 queue using the MBH algorithm takes on an HP700 workstation less than 2 seconds.

Remark: For a constant service time distribution D with $\Pr(H = 1)$, we have $\gamma_1 = 1$. A typical example for such a service time distribution is the statistical ATM multiplexer.

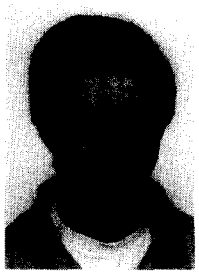
References

[1] A. Baiocchi, N. Belfari Melazzi, M. Listanti, A. Roveri and R. Winkler, Loss performance analysis of an ATM multiplexer loaded with high-speed ON-OFF sources, *IEEE J. Sel. Areas Comm.* **9** (3) (1991) 388–393.
[2] C. Blondia, Finite capacity vacation models with non-renewal input, *J. Appl. Probab.* **28** (1991) 174–197.
[3] C. Blondia and O. Casals, Performance analysis of statistical multiplexing of VBR sources, *Proc. IEEE Infocom*, May 1992, pp. 828–838.
[4] U. Briem, T.H. Theimer and H. Kroner, A general discrete-time queueing model: analysis and applications, in: A. Jensen and V.B. Iversen (Eds.), *Teletraffic and Datatraffic in a Period of Change*, North-Holland Studies in Telecommunication **14**, North-Holland, Amsterdam (1991) 13–19.
[5] S.C. Dafermos and M.F. Neuts, A single server queue in discrete time, *Cah. Cent. Etude Rech. Op.* **13** (1971) 23–40.
[6] W. Ding, A unified correlated input process model for telecommunication networks, in: A. Jensen and V.B. Iversen (Eds.), *Teletraffic and Datatraffic in a Period of Change*, North-Holland Studies in Telecommunication **14**, North-Holland, Amsterdam (1991) 539–544.
[7] W. Ding and P. Decker, Waiting time distribution of a discrete SSMP/G/1 queue and its implications in ATM systems, *7th ITC Specialist Seminar*, Morristown, October 1990, paper 9.4.
[8] S. Fuhrmann and J.Y. Le Boudec, Burst and cell level models for ATM buffers, in: A. Jensen and V.B. Iversen (Eds.) *Teletraffic and Datatraffic in a Period of Change*, North-Holland Studies in Telecommunication **14**, North-Holland, Amsterdam (1991) 975–980.
[9] B. Fontana and A. Guerrero, Packet traffic characterization, arrival laws and waiting times, *Proc. 12th Int. Teletraffic Congress*, Torino, June 1988, paper 4.2A.4.
[10] W.A. Gardner, *Introduction to Random Processes*, Macmillan (1986).
[11] D. Gross and D.M. Harris, *Fundamentals of Queueing Theory*, 2nd ed., Wiley, New York (1985).
[12] R. Grunenfelder and L. Zubieta, Measurement of ATM traffic on the cell, burst and activity level by traffic sampling, *Proc. IEEE Infocom*, Florence, May 1992, pp. 479–486.
[13] O. Hashida, Y. Takahashi and S. Shimogawa, Switched batch Bernoulli process (SBBP) and the discrete-time SBBP/G/1 queue with application to statistical multiplexer performance, *IEEE J. Sel. Areas Comm.* **9** (3) (1991) 394–401.
[14] H. Heffes and D.M. Lucantoni, A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE J. Sel. Areas Comm.* **4** (6) (1986) 856–868.
[15] C. Herrmann, Analysis of the discrete-time SMP/D/1/s finite buffer queue with applications in ATM, *Proc. IEEE Infocom '93*, San Francisco, April 1993, pp. 160–167.
[16] CCITT Recommendation I. 35B, B-ISDN ATM Layer Cell Transfer Performance, Geneva, 1993.
[17] CCITT Recommendation I.363, B-ISDN ATM Adaptation Layer (AAL) Specification, Geneva, 1993.

[18] CCITT Recommendation I.371, Traffic Control and Congestion Control in B-ISDN, Geneva, 1993.
[19] J.-Y. Le Boudec, An efficient solution method for Markov models of ATM links with loss priorities, *IEEE J. Sel. Areas Comm.* **9** (3) (1991) 408–417.
[20] P. Levy, Processus semi-markoviens, *Proc. Int. Cong. Math*, Amsterdam, 1954, Vol. 3, pp. 416–426.
[21] M. Luoni, ATM traffic characterization with applications to connection acceptance control, Thèse No. 979, Ecole Polytechnique Fédérale Lausanne, 1991.
[22] K.S. Meier-Hellstern, A fitting algorithm for Markov-modulated Poisson processes having two arrival rates, *Eur. J. Op. Res.* **29** (1987) 370–377.
[23] M.F. Neuts, *Matrix Geometric Solutions in Stochastic Models*, John Hopkins University Press, Baltimore (1981).
[24] R. Pyke, Markov renewal processes: definitions and preliminary properties, *Ann. Math. Statist* **32** (1961) 1231–1242.
[25] M.H. Rossiter, A switched Poisson model for data traffic, *Aust. Telecomm. Res.* **21** (1987) 53–57.



Reto Grünenfelder received the diploma degree in physics from ETH Zurich in 1986 and the Ph.D. in electrical engineering from EPFL Lausanne in 1991.
From 1986 to 1987, he was with Siemens Albis AG, Zurich, where he mainly worked in the fields of microwave engineering and signal processing. From 1987 to 1992, he was with EPFL, researching in the area of stochastic modelling and communication systems and networks. He has been involved in the RACE projects R1022 "Technology for ATD" and R2032 COMBINE (COMposite Broadband INterworking and End-to-end models). In summer 1992 he joined the technology management of Alcatel STR in Zurich where he mainly works in the fields of broadband testing, performance evaluation of communication networks and technology marketing.



Stephan Robert was born in Neuchâtel, Switzerland, in 1965. He graduated as an engineer in microengineering from the Swiss College of Engineering (ETS), Le Locle (Switzerland), in 1986. He received the Dipl. Ing. degree in electrical engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne (Switzerland), in 1991.
From 1986 to 1987, he was employed at Asea Brown Boveri (ABB), Baden (Switzerland), where he worked in the field of electronics. From 1991 to 1992, he was employed at EPFL in the Mechanical Engineering Department where he worked in the field of atmospheric turbulences and signal processing. In 1992, he joined the Electrical Engineering Department where he is working towards the Ph.D. degree. His research include performance modelling, queueing theory and analysis of broadband networks.