Exact Results for an ATM Multiplexer with Infinite Queue Loaded with Batch Markovian Arrivals*

Rasti Slosiar [a]

[a]Swiss Federal Institute of Technology (EPFL), Electrical Engineering Department (DE), Telecommunications Laboratory, CH-1015 Lausanne, Switzerland

Abstract

Important efforts have been invested to characterize the performance of an ATM multiplexer system loaded with a set of superimposed ON/OFF sources, since this ON/OFF source model has been extensively used as a coarse approximation of variable bitrate services (e.g. shaped LAN traffic). Literature exists on approximations for queueing systems, in which a large number of superimposed sources is considered, and replaced by a more tractable aggregate model. In this contribution, the focus is set on a discrete-time technique related to exact results for the ON-OFF$^{[N]}$/D/1 queueing system with infinite queueing capacity for the superposition of a small number of such sources. The discrete-time batch Markovian arrival process (D-BMAP) *Queue Input Process* (QIP) is first introduced. Complete solutions for the D-BMAP/D/1 queueing system with normalised service time are provided in terms of the system occupancy distribution, its moments, the sojourn time distribution and its moments. A proposal for expressing an aggregate of ON/OFF sources QIP as a D-BMAP is then formulated, and the sparsity of the matrices and vectors highlighted. Finally, the degenerate cases corresponding to no multiplexing gain ("underload" case), the superposition of two-state Markov sources (2SM) and Bernoulli sources are handled, and closed form expressions for the moments provided.

Keywords: ATM, D-BMAP, Queueing Theory

1. INTRODUCTION

Lots of efforts have been invested in finding techniques to determine the loss probability and delay of a group of homogeneous ON/OFF sources entering an ATM multiplexer of finite capacity. As the state space becomes extremely large as the parameters of the multiplexer system grows, exact analytical methods have been left aside in profit of less involving approximations methods. Approximate solutions have been obtained by relaxing some conditions in the aggregate of ON/OFF sources, e.g. by its replacement by a *Markov Modulated Poisson Process* (MMPP) source model [1].

In this contribution, we formulate the exact solution of an infinite-size queueing system (multiplexer) loaded by a bunch of N homogeneous discrete-time three-parameter ON/OFF

*This research was supported by the Swiss Federal Office of Education and Science (OFES) within a funding for project RACE 2024 entitled "Broadband Access Facilities".

D. D. Kouvatsos (ed.), *Performance Modelling and Evaluation of ATM Networks*

sources. The queueing system can be completely solved using discrete-time matrix-analytic techniques, when the aggregate traffic is expressed as one single *Queue Input Process* (QIP): the *Discrete-time Batch Markovian Arrival Process* (D–BMAP) [2]. The limits of such methods are of course the high state space involved. However, the usage of a clever D–BMAP expression of the aggregated sources and some sparse matrix/vector computation software yields the solution of queueing systems loaded by more than just a few ON/OFF sources. The results will be further used in future work as a basis for comparison with the performance of distributed access networks where the traffic scheduling may be operated by a Medium Access Control (MAC) Protocol, where the number of test sources is low anyway.

In section 2, we present solution techniques of the general D-BMAP/D/1 queue, its moments, the sojourn time (waiting and service time) statistics and its moments. These methods are then applied in section 3 to the superposition of a homogeneous set of ON/OFF sources, to "overload" (with multiplexing gain) and "underload" (no multiplexing gain) cases (section 3.3). Results for degenerated cases (two-state Markov and Bernoulli sources) are handled in section 4. Finally, section 5 illustrates the numerous developments by some numerical examples.

2. GENERAL SOLUTION OF A D-BMAP/D/1 QUEUE

A good overview of the techniques for solving the continuous-time BMAP/G/1 can be found in [7] and [8]. In [2], the authors introduce an algorithm for assessing the ATM multiplexer performance through the analysis of the D–BMAP/D/1/K (finite capacity) queue. A recursive algorithm is presented for the computation of the loss probability for increasing multiplexer queue sizes. The D–MAP/G/1 (no batch arrivals) has been handled in [3].

2.1. The D–BMAP and Multiplexer Models

We assume that at each time slot, a random number of arrivals between 0 and N may occur at the queue input, which corresponds to a maximum of one arrival on each inlet of the multiplexer model (location ① on fig. 1). The D–BMAP models the batch arrival process after aggregation ②, this process constitutes the *Queue Input Process* (QIP). The D-BMAP has already been described in [2] : a discrete-time *Markov Chain* (MC) governs the cell generation process. We assume that the governing MC is irreducible and we define by H its number of states. The probability that the QIP MC produces a batch of n arrivals , $0 \le n \le N$, while transiting from state i to state j is given by the matrix element $[\boldsymbol{D}_n]_{i,j}$, $1 \le i,j \le H$ (the operator $[\cdot]_{i,j}$ returns the element in row i and column j). Consequently, matrix $\boldsymbol{D}_n$ encompasses all transitions occurring in the QIP MC accompanied by the generation of n simultaneous cell arrivals. The underlying source MC transition matrix $\boldsymbol{D}$ of dimensions $H \times H$ is then the sum of the $(N+1)$ submatrices $\boldsymbol{D}_n$, $0 \le n \le N$:

$$\boldsymbol{D} = \sum_{n=0}^{N} \boldsymbol{D}_n \tag{1}$$

The ATM multiplexer is modeled as depicted in figure 1; a set of N inlets, numbered from 1 to N deliver a maximum of one cell per ATM slot to the multiplexer ①. Therefore, a maximum of N aggregated cells may arrive within one time slot into the queueing system ②. The ordering of the cells within the multiplexing function is supposed to be performed randomly, i.e. if a batch of $a_k > 0$ cells arrive at time slot k, the probability for a tagged cell within this batch to be placed as i-th cell into the queue is independent of i and equals $1/a_k$.

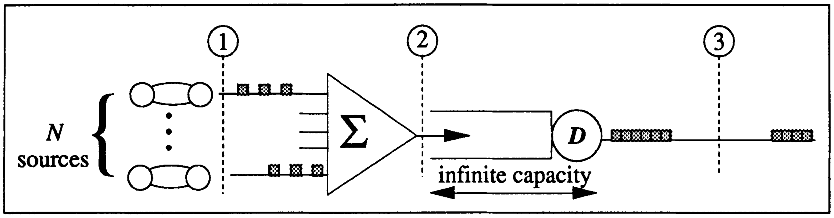

Figure 1: Multiplexer model.

The per-cell service time is constant and defines the unit of time; it corresponds to the cell transmission duration. The queue size n_k at time slot k is therefore governed by the following recursive expression:

$$n_k = [n_{k-1} - 1]^+ + a_k, \quad n_k \geq 0, \quad 0 \leq a_k \leq N \tag{2}$$

where a_k is the number of cells in the batch arrival at time slot k described by the QIP MC matrix $\boldsymbol{D}$, and $[\cdot]^+$ is defined as max[0,·]. Under these circumstances, the aggregate offered traffic ρ to the queueing system can be computed as [2]:

$$\rho = N\lambda = \vec{d} \cdot \left(\sum_{n=1}^{N} n \boldsymbol{D}_n \right) \cdot \vec{e}_H^{\mathrm{T}} \tag{3}$$

where $\vec{e}_H$ is a row vector of dimension 1xH consisting of ones, $[\cdot]^{\mathrm{T}}$ denotes the transpose operation, $\vec{d}$ is the steady state row vector of the arrival process MC defined by:

$$\vec{d} \cdot \boldsymbol{D} = \vec{d} \tag{4}$$

together with

$$\vec{d} \cdot \vec{e}_H^{\mathrm{T}} = 1, \tag{5}$$

and λ is the per-inlet ① offered traffic. At any time slot, the state of the system under study can be identified with the couple of *Random Variables* (RVs) $\{Q, X\}$, where $Q \geq 0$ describes the number of cells in the system (queue and server), and $1 \leq X \leq H$ the D–BMAP state at the same point in time. The multiplexer can be modeled by the bi-dimensional semi-Markov chain of transition matrix $\boldsymbol{M}$:

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{D}_0 & \boldsymbol{D}_1 & \cdots & \boldsymbol{D}_N & \boldsymbol{0} & \cdots & & \\ \boldsymbol{D}_0 & \boldsymbol{D}_1 & \cdots & \boldsymbol{D}_N & \boldsymbol{0} & \cdots & & \\ \boldsymbol{0} & \boldsymbol{D}_0 & \boldsymbol{D}_1 & \cdots & \boldsymbol{D}_N & \boldsymbol{0} & \cdots & \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}_0 & \boldsymbol{D}_1 & \cdots & \boldsymbol{D}_N & \boldsymbol{0} & \cdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \ddots & \ddots \\ & & & & & & & \end{bmatrix} \tag{6}$$

2.2. Formulation of the Queueing Equation

The probabilistic formulation of the queueing system corresponding to the time-domain multiplexer equation (2) can be performed in *Matrix Probability Generating Function* (MPGF) form. A probability generating function (PGF) $\tilde{X}(z)$ for a random variable X is defined as:

$$\tilde{X}(z) = \sum_{i=0}^{\infty} \mathrm{P}[X = i]\, z^i \tag{7}$$

where P[·] denotes the probability of an event. Similarly, we define the MPGF of the arrival process as the matrix $\boldsymbol{D}(z)$ whose elements $[\boldsymbol{D}(z)]_{i,j}$ represent partial PGFs of the number of arrivals when a transition from state i to state j occurs:

$$\boldsymbol{D}(z) = \sum_{n=0}^{N} \boldsymbol{D}_n z^n \tag{8}$$

The per-slot average number of arrivals (corresponding to the offered traffic since the service time is deterministic and set to "one time slot") can be derived from (8) as well, by derivation with respect to z:

$$\vec{d} \cdot \boldsymbol{D}'(1) \cdot \vec{e}_H^{\,\mathrm{T}} = \vec{d} \cdot \frac{\mathrm{d}}{\mathrm{d}z} \boldsymbol{D}(z) \bigg|_{z=1} \cdot \vec{e}_H^{\,\mathrm{T}} = \vec{d} \cdot \sum_{n=1}^{N} n\boldsymbol{D}_n \cdot \vec{e}_H^{\,\mathrm{T}} = \rho \tag{9}$$

We further define the vector PGF of the queue size distribution $\vec{q}(z)$ whose elements give the queue size partial PGF when the D–BMAP source process is in a particular state as:

$$\vec{q}(z) = \sum_{l=0}^{\infty} \vec{q}_l\, z^l \tag{10}$$

where element j of the row vector $\vec{q}_l$ gives the joint probability that the D–BMAP is in phase j while the queue size is l:

$$[\vec{q}_l]_j = \mathrm{P}[Q = l,\, X = j] \tag{11}$$

In order to express the nonlinear multiplexer equation (2) and remove the nonlinear operator $[\cdot]^+$, the boundary joint probability values corresponding to an empty system need to be introduced:

$$[\vec{q}_0]_j = \mathrm{P}[Q = 0,\, X = j] \tag{12}$$

From the elements defined in (8)–(12), we can formulate the multiplexer equation (2) in MPGF form as follows :

$$\vec{q}(z) = \left[\frac{1}{z} \vec{q}(z) + \frac{z-1}{z} \vec{q}_0 \right] \cdot \boldsymbol{D}(z) \tag{13}$$

Expression (13) shows basically that the queue size ***Probability Mass Function*** (PMF):

$$q(l) = \mathrm{P}[Q = l] \tag{14}$$

equals the convolution of the shifted-folded queue size PMF and the number of arrivals PMF. The "folding" operation consists in removing the probabilities of negative values of a RV (i.e. $\mathrm{P}[Q < 0]$) and adding them to the probability of value 0 ($\mathrm{P}[Q = 0]$). Extraction of the queue vector $\vec{q}(z)$ PGF from (13) finally yields :

$$\vec{q}(z) = (z-1)\,\vec{q}_0 \cdot \boldsymbol{D}(z) \cdot \left[z\boldsymbol{I}_H - \boldsymbol{D}(z)\right]^{-1} \tag{15}$$

where $\boldsymbol{I}_H$ denotes the identity matrix of dimensions $H{\times}H$. The probability $q(l)$ that the queue size equals l, and the probability generating function (PGF) $\tilde{Q}(z)$ of the queue distribution are immediately available from the previously defined vectors :

$$\begin{aligned} q(l) &= \vec{q}_l \cdot \vec{e}_H^{\,\mathrm{T}} \\ \tilde{Q}(z) &= \vec{q}(z) \cdot \vec{e}_H^{\,\mathrm{T}} \end{aligned} \tag{16}$$

An essential and challenging task consists in finding the boundary probability vector $\vec{q}_0$ from which subsequent queueing performance variables can be derived.

2.3. Known Approaches for the Determination of the Boundary Values Vector

From the boundary values vector $\vec{q}_0$ that appeared in the queueing equation (15), all the performance variables of the queueing system with batch Markovian arrivals can be characterized. This boundary vector is well-known to play an essential role in the queueing theory, and various approaches have been applied in order to obtain this vector. For the exact resolution of queueing systems with correlated arrivals, at least three approaches are well-known:

- the iterative matrix-analytic solution based on the "first passage" transition matrix $\boldsymbol{G}$, [10], [7];
- the eigensystem decomposition of the queueing system equation (15), [6], [17].
- the functional-equation approach [4], [16].

We first sketch the principles of an eigenvalue decomposition and the equations that reveal the boundary probability values. An eigenvalue decomposition of the system equation MPGF (15) reads:

$$\vec{q}(z) = (z-1)\vec{q}_0 \cdot \sum_{i=1}^{H} \frac{\lambda_i(z)}{z - \lambda_i(z)} \vec{v}_{R,i}^{\,\mathrm{T}}(z) \cdot \vec{v}_{L,i}(z) \tag{17}$$

where $\lambda_i(z)$ are the polynomial eigenvalues of $\boldsymbol{D}(z)$, and $\vec{v}_{R,i}^{\,\mathrm{T}}(z)$, $\vec{v}_{L,i}(z)$ are its associated right and left polynomial eigenvectors respectively. See [6] for a discussion of the number and existence of the eigenvalues. The set of roots composed by the zeroes in $z{-}\lambda_i(z) = 0$ for each

i, contains the poles that make part of the queueing system PGF $\tilde{Q}(z)$. Within this set, the roots that lie inside the unit circle have to vanish from the solution, since we assumed that the system operates in a stable configuration $\rho < 1$. Therefore, the equation:

$$\vec{q}_0 \cdot \vec{v}_{R,i}^{\mathrm{T}}(z) = 0 \tag{18}$$

must be respected at each vanishing root $z = z_k$. These equations, with some others, yield the boundary values in $\vec{q}_0$.

This elegant approach attempting to express the queue size PGF $\vec{q}(z)$ via the eigenvalue decomposition of the MPGF (15) has revealed to be effective only when the polynomial eigenvalues, $\lambda_i(z)$, of $\boldsymbol{D}(z)$ can be expressed analytically so that the vanishing and non-vanishing poles in $\boldsymbol{D}(z)\cdot[z\boldsymbol{I}-\boldsymbol{D}(z)]^{-1}$ can be computed numerically, [6], [17]. This is obviously possible only when the dimension of the traffic matrix $\boldsymbol{D}(z)$, i.e. H, is below 5, or when the system is a superposition of sources which are governed by a MC of dimension smaller than 5.

The last approach among the three we mentioned earlier is called the functional-equation technique and has been applied in [4] for the computation of moments, while it has been tailored in [16] for the computation of the tail of queue distributions. This approach does not reveal the boundary values, but provides a handy relationship between some queueing system performance variables and the boundary values once they are known.

We have applied this method in [12] for the derivation of closed form expressions of the moments of the distributions of the number of customers in a queueing system, when it is loaded with a homogeneous set of i.i.d. three-parameter discrete-time ON/OFF sources. These expressions are function of the source parameters and the boundary probabilities.

For what concerns the discrete-time ON/OFF source model and its multiple superposition (see section 3), the eigensystem decomposition appears to be precarious due to the high dimension of the D-BMAP matrices involved (often higher than 4) and we shall therefore apply the matrix-analytic solution throughout this work.

2.4. Boundary Values Determination Based on the *G* Matrix

In this section, we shall introduce the "first-passage" matrix $\boldsymbol{G}$ and how an infinite–capacity queueing system loaded with batch-Markovian arrivals may be entirely solved by the determination of this key matrix.

2.4.1. The "First Passage" Matrix G

The accuracy of the solution to our D–BMAP/D/1 queueing system resides mainly in the accuracy of the computation of the "first passage" matrix $\boldsymbol{G}$ ([10], [7]). The elements of the matrix $\boldsymbol{G}$ describe the change of phase in the source process when the queue returns for the first time from level l to level l–1. More formally, $[G]_{i\,j}$ express the conditional probability that the D–BMAP enters state j when the number of cells in the system gets for the first time to l–1(after any number of transitions ≥ 1), given that it started from state i in the D-BMAP when the number of cells was l, $l \geq 1$.

The $\boldsymbol{G}$ matrix plays a crucial role in the queueing system theory since it is not dependent on the initial queue size l [10]. The $\boldsymbol{G}$ matrix is found to be stochastic (see [10, p. 88] for a discussion of the $\boldsymbol{G}$ matrix) and satisfies the following non-linear equation:

$$G = \sum_{n=0}^{N} D_n \cdot G^n \tag{19}$$

which can be solved iteratively according to

$$G(k+1) = \sum_{n=0}^{N} D_n \cdot [G(k)]^n \tag{20}$$

It has been advised to compute the summation in (19) using the recursive Horner's algorithm [10, p.162]. Horner's algorithm consists in starting with the matrix:

$$Y_0 = D_N \tag{21}$$

and then compute recursively each Y_j as follows up to Y_N, which yields the final result corresponding to the right hand side of equation (19):

$$Y_j = D_{N-j} + Y_{j-1} \cdot G, \quad 1 \le j \le N \tag{22}$$

The start of the iterative computation (20) could be a stochastic matrix e.g. $G(0) = D$, which reveals to speed up the convergence. Intuitively, expression (19) tells us that the "first passage" from queue size l to queue size l–1 is the same as the one obtained when any n arrivals occur followed by a decrease of queue size by n cells. We see also that the transitions given by their non-null probabilities in D_0 are the only ones in the D–BMAP that allow the system size to be reduced by one within one transition.

2.4.2. Boundary Values

Once the relationships that tie successive queue length vectors (11) together via the G matrix are known, the determination of the boundary values and of the system occupancy distribution is straightforward. As we shall see, the boundary probability vector $\vec{q}_0$ is in fact directly related to the first passage matrix left eigenvector $\vec{g}$ associated with the eigenvalue 1, i.e. the steady state vector of the underlying MC, defined by $\vec{g} \cdot G = \vec{g}$ and $\vec{g} \cdot \vec{e}_H^{\,\mathrm{T}} = 1$, which stresses the importance of the "first passage" matrix.

In order to determine $\vec{q}_0$, we first remember that the "first passage" matrix G encompasses all the possible transitions "above" a given level l until the system reaches the level l–1 just below l for the first time. From the previous assumption, an equation for the determination of $\vec{q}_0$ can be drawn as follows:

1. we start with an ATM multiplexer initially empty,
2. a number of arrivals i occurs in the time slot following the one where the system was empty, with $0 \le i \le N$,
3. the system returns to an empty state by stepping downwards i levels; this decrease corresponds to i "first passages".

In a steady state situation, the probability of the system to be empty before step 1 is the same as the probability after step 3. Consequently, the intuitive formulation of steps 1–3 enumerated above can be expressed analytically as follows:

$$\vec{q}_0 \cdot \sum_{n=0}^{N} \boldsymbol{D}_n \cdot \boldsymbol{G}^n = \vec{q}_0 \tag{23}$$

In fact, in (23), the matrix $\sum_{n=0}^{N} \boldsymbol{D}_n \cdot \boldsymbol{G}^n$ expresses the phase transitions in the D–BMAP between two successive epochs where the system is empty, and encompasses the complete system evolution between such epochs. The steady state vector $\vec{g}$ of the underlying MC provides the stationary phase vector of the D–BMAP at instants where the system is empty.

Observing that the matrix $\sum_{n=0}^{N} \boldsymbol{D}_n \cdot \boldsymbol{G}^n$ is nothing else than $\boldsymbol{G}$ itself according to (19), we deduce that $\vec{q}_0$ is an invariant vector of $\boldsymbol{G}$. Moreover, as the probability that the system is empty is given by 1–ρ, we obtain directly:

$$\vec{q}_0 = (1-\rho)\vec{g} \tag{24}$$

where $\vec{g}$ is the steady state vector of $\boldsymbol{G}$, as defined earlier, and $q(0) = \vec{q}_0 \cdot \vec{e}_H^{\,\mathrm{T}} = 1 - \rho$.

2.5. The Queue Distribution

Using the "first passage" matrix again, the complete system occupancy distribution can be characterized similarly to the case of the D–MAP/G/1 queue [3]. The steps to be followed are similar to the ones in the previous section. We aim at expressing the relationships that lead to a particular level l, from all other candidate levels using the arrival process matrices $\boldsymbol{D}_i$, and the matrix $\boldsymbol{G}$ corresponding to a single level decrease. We will conduct our reasoning under the assumption (without loss of generality) that level l is far enough from level 0, i.e. that the considered level l is higher than N. Levels $1 \le l \le N$ can be obtained similarly by adjusting the summation bounds. We observe the following (see figure 2) :

- The lowest starting level from which the target level $l > N$ can be reached within one time slot is l–N+1, since at most N arrivals may occur within one time slot and one cell departs.
- The highest starting level is bounded by l, since levels higher than l are covered by the "first passage" matrix. A possible starting level j is then bounded as $l\text{-}N+1 \le j \le l$.
- When starting with a particular level j as bounded above, at least l-j+1 arrivals need to occur in the first time slot, in order to reach at least level l. The number of arrivals n must be therefore in the range $l\text{-}j+1 \le n \le N$.
- After n arrivals occurred, the system level will have reached j+n-1 (n arrivals and one cell departure), which is higher or equal to the target level l.
- Exactly j+n-1-l "first passages" must occur in order to bring the system back to level l.

An equation tying all possible starting levels to the destination level can be therefore expressed:

$$\vec{q}_l = \sum_{j=l-N+1}^{l} \vec{q}_j \sum_{n=l-j+1}^{N} \boldsymbol{D}_n \cdot \boldsymbol{G}^{j+n-1-l}, \quad l > N \tag{25}$$

Extracting $\vec{q}_l$ from (25) directly yields the probability vector for level l as a function of the N-1 previous levels (29). However, some care has to be applied for levels between 1 and N, since:

- level 0 becomes a possible starting level, and due to the nonlinear operator $[\cdot]^+$ in (2), the intermediate level becomes $j+n$ instead of $j+n-1$, and the powers of the matrix $\boldsymbol{G}$ in (25) have to be adjusted differently;
- the summation needs to be performed from level one instead of level $l-N+1 \leq 1$.

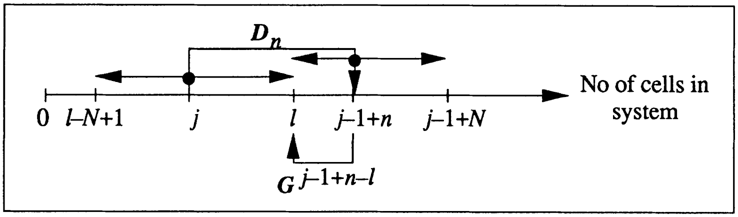

Figure 2: Recursive relationship between successive queue lengths.

Following these precautions, a complete set of equations for the determination of the complete queue occupancy PMF can be obtained from (16), from the boundary probability vector, from the $\boldsymbol{G}$ matrix and from the D-BMAP. In order to simplify the computations, let us define the intermediate matrices $\boldsymbol{V}(l)$:

$$\boldsymbol{V}(l) = \sum_{n=l}^{N} \boldsymbol{D}_n \cdot \boldsymbol{G}^{n-l}, \quad 1 \leq l \leq N \tag{26}$$

The expressions for the joint probability vectors of the number of cells in the system and the state of the D-BMAP can be then computed as:

$$\vec{q}_1 = [\vec{q}_0 \cdot \boldsymbol{V}(1)] \cdot \boldsymbol{W}, \qquad l = 1 \tag{27}$$

$$\vec{q}_l = \left[\vec{q}_0 \cdot \boldsymbol{V}(l) + \sum_{j=1}^{l-1} \vec{q}_j \cdot \boldsymbol{V}(l-j+1)\right] \cdot \boldsymbol{W}, \qquad 1 < l \leq N \tag{28}$$

$$\vec{q}_l = \left[\sum_{j=l-N+1}^{l-1} \vec{q}_j \cdot \boldsymbol{V}(l-j+1)\right] \cdot \boldsymbol{W}, \qquad N < l \tag{29}$$

where the matrix $\boldsymbol{W}$ is defined as :

$$\boldsymbol{W} = [\boldsymbol{I}_H - \boldsymbol{V}(1)]^{-1} \tag{30}$$

We observe that during the computation of the queue size PMF from (27)-(29), at most N successive queue vectors need to be stored, including the currently computed level.

2.6. Moments of the Queue Distribution

In this section, we highlight the algorithm for the derivation of the moments of the queue occupancy distribution, and focus on the first two moments whose expressions are explicitly provided. The first two moments $Q^{[1]} = \mathrm{E}[Q]$ and $Q^{[2]} = \mathrm{E}[Q^2] = \mathrm{Var}[Q]+\mathrm{E}[Q]^2$ of the queueing system can be computed from the factorial moments obtained by derivation of the queue length PGF (16) and by subsequent setting of z=1:

$$\begin{aligned} Q^{[1]} &= \sum_{l=1}^{\infty} l\, q(l) = \frac{\mathrm{d}}{\mathrm{d}z}\tilde{Q}(z)\bigg|_{z=1} \\ Q^{[2]} &= \sum_{l=1}^{\infty} l^2\, q(l) = \frac{\mathrm{d}^2}{\mathrm{d}z^2}\tilde{Q}(z)\bigg|_{z=1} + Q^{[1]} \end{aligned} \tag{31}$$

The moments of the D–BMAP/D/1 queueing system can be computed starting from (15), by derivation with respect to z. Let us call $\boldsymbol{D}'(1)$, $\boldsymbol{D}''(1)$, and $\boldsymbol{D}'''(1)$ the three first factorial moments of the matrix $\boldsymbol{D}(z)$ obtained by successive derivation of $\boldsymbol{D}(z)$ with respect to z and by subsequent setting z=1:

$$\boldsymbol{D}'(1) = \sum_{i=1}^{N} i\boldsymbol{D}_i, \qquad \boldsymbol{D}''(1) = \sum_{i=2}^{N} i(i-1)\boldsymbol{D}_i, \qquad \boldsymbol{D}'''(1) = \sum_{i=3}^{N} i(i-1)(i-2)\boldsymbol{D}_i \tag{32}$$

Of course, when N is smaller than 3, some of the previous derivations lead to null matrices. Following the steps in [10, p.143], the moments of the queue distribution can be obtained by derivation of (15) with respect to z. The approach is illustrated hereafter for the first moment of the system occupancy distribution, higher moments can be obtained recursively following the same steps. The determination of the first moment $Q^{[1]} = \vec{q}\,'(1)\cdot\vec{e}_H^{\mathrm{T}}$ starts from the first (33) and second (34) derivatives of both sides of expression (15) with subsequent setting z=1:

$$\vec{q}\,'(1)\cdot\left[\boldsymbol{I}_H - \boldsymbol{D}\right] = \vec{q}_0\cdot\boldsymbol{D} + \vec{d}\cdot\left[\boldsymbol{D}'(1) - \boldsymbol{I}_H\right] \tag{33}$$

$$\vec{q}\,''(1)\cdot\left[\boldsymbol{I}_H - \boldsymbol{D}\right] + 2\vec{q}\,'(1) = 2\left[\vec{q}_0 + \vec{q}\,'(1)\right]\cdot\boldsymbol{D}'(1) + \vec{d}\cdot\boldsymbol{D}''(1) \tag{34}$$

since we observe that $\vec{q}(z)\big|_{z=1} = \vec{d}$ and $\boldsymbol{D}(z)\big|_{z=1} = \boldsymbol{D}$, where $\boldsymbol{D}$ was defined in (1). While aiming at the extraction of $q'(1)$ in (33), we observe also that $\boldsymbol{I}_H - \boldsymbol{D}$ is not inversible because $\boldsymbol{D}$ is stochastic and the value "one" is part of its eigenvalues; however, by adding $\left[\vec{q}\,'(1)\cdot\vec{e}_H^{\mathrm{T}}\right]\vec{d} = Q^{[1]}\vec{d}$ to both sides of the expression, we obtain the inversible matrix $\boldsymbol{Z}^{-1} = \boldsymbol{I}_H - \boldsymbol{D} + \vec{e}_H^{\mathrm{T}}\cdot\vec{d}$ and finally $q'(1)$ reads:

$$\vec{q}\,'(1) = Q^{[1]}\vec{d} + \left[\vec{q}_0\cdot\boldsymbol{D} + \vec{d}\cdot\left[\boldsymbol{D}'(1) - \boldsymbol{I}_H\right]\right]\cdot\boldsymbol{Z} \tag{35}$$

Since we still have two unknowns ($Q^{[1]}$ and $\vec{q}\,'(1)$), we need a second equation in order to eliminate $q'(1)$. We achieve this by postmultiplying (34) by $\vec{e}_H^{\mathrm{T}}$ and observing that

$(\boldsymbol{I}_H - \boldsymbol{D})\cdot \vec{e}_H^{\mathrm{T}} = \vec{0}^{\mathrm{T}}$. These operations yield:

$$2Q^{[1]} = \left[2[\vec{q}_0 + \vec{q}\,'(1)]\cdot \boldsymbol{D}'(1) + \vec{d}\cdot \boldsymbol{D}''(1)\right]\cdot \vec{e}_H^{\mathrm{T}} \tag{36}$$

Inserting (35) into (36) and taking into account that $Q^{[1]}\vec{d}\cdot \boldsymbol{D}'(1)\cdot \vec{e}_H^{\mathrm{T}} = Q^{[1]}\rho$ (cf. expr. (9)), we may now extract $Q^{[1]}$ and obtain finally:

$$Q^{[1]} = \frac{1}{2(1-\rho)}\left\{2\left[\left(\vec{q}_0\cdot \boldsymbol{D} + \vec{d}\cdot[\boldsymbol{D}'(1) - \boldsymbol{I}_H]\right)\cdot \boldsymbol{Z} + \vec{q}_0\right]\cdot \boldsymbol{D}'(1) + \vec{d}\cdot \boldsymbol{D}''(1)\right\}\cdot \vec{e}_H^{\mathrm{T}} \tag{37}$$

where

$$\boldsymbol{Z} = \left[\boldsymbol{I}_H - \boldsymbol{D} + \vec{e}_H^{\mathrm{T}}\cdot \vec{d}\right]^{-1} \tag{38}$$

Similar steps provide the second moment of the queue occupancy distribution :

$$\begin{aligned} Q^{[2]} = Q^{[1]} + \frac{1}{3(1-\rho)}\Big\{&3\left(2\vec{q}_0\cdot \boldsymbol{D}'(1) + 2\vec{q}\,'(1)\cdot[\boldsymbol{D}'(1) - \boldsymbol{I}_H] + \vec{d}\cdot \boldsymbol{D}''(1)\right)\cdot \boldsymbol{Z}\cdot \boldsymbol{D}'(1) \\ &+ 3[\vec{q}_0 + \vec{q}\,'(1)]\cdot \boldsymbol{D}''(1) + \vec{d}\cdot \boldsymbol{D}'''(1)\Big\}\cdot \vec{e}_H^{\mathrm{T}} \end{aligned} \tag{39}$$

Higher moments (say of order n) can be obtained recursively starting with the n-th and (n+1)-th derivative of (15), while re-using results at the previous levels. In the case of a superposition of a set of i.i.d. ON/OFF sources, a representation based on the source parameters which minimizes the vector-matrix operations can be found thank to the functional-equation technique initially developed in [4], these results can be found in [12].

2.7. The Sojourn Time Distribution

Once the joint probabilities $[\vec{q}_l]_j = \mathrm{P}[Q = l, X = j]$ are known, the sojourn time W PMF, $w(n) = \mathrm{P}[W = n]$, of any cell entering the multiplexer, can be also derived. We have defined the sojourn time as the sum of the waiting time in the queue and of the service time (which remains constant and equals the time unit in our case) of the equivalent queueing system. The probability that an arbitrary cell waits n time slots in the multiplexer before leaving it equals the probability that the cell encounters a system already filled with n–1 cells when it enters the system, which sums up to n sojourn time slots when taking into account its own service time. This assumption implies also that the cells are served according to a *First-In-First-Out* (FIFO) service policy. The observation of the system contents is made after a possible departure: $n'_{k+1} = [n_k - 1]^+$ (compare with (2)). Taking these assumptions into account, the sojourn time PMF can be computed as follows:

$$w(1) = \frac{1}{\rho}(\vec{q}_0 + \vec{q}_1)\cdot[\boldsymbol{D} - \boldsymbol{D}_0]\cdot \vec{e}_H^{\mathrm{T}} = \frac{1}{\rho}(\vec{q}_0 + \vec{q}_1)\cdot\left[\vec{e}_H^{\mathrm{T}} - \boldsymbol{D}_0\cdot \vec{e}_H^{\mathrm{T}}\right], \qquad n = 1 \tag{40}$$

$$w(n)=\frac{1}{\rho}\vec{q}_0\cdot\left[\sum_{i=n}^{N}\boldsymbol{D}_i\right]\cdot\vec{e}_H^{\mathrm{T}}+\frac{1}{\rho}\sum_{l=1}^{n}\vec{q}_l\cdot\left[\sum_{j=n-l+1}^{N}\boldsymbol{D}_j\right]\cdot\vec{e}_H^{\mathrm{T}},\qquad 1<n\le N \tag{41}$$

$$w(n)=\frac{1}{\rho}\sum_{l=n-N+1}^{n}\vec{q}_l\cdot\left[\sum_{i=n-l+1}^{N}\boldsymbol{D}_i\right]\cdot\vec{e}_H^{\mathrm{T}},\qquad N<n \tag{42}$$

Note that a cell sojourns at least during one time slot in the multiplexer. The factor $1/\rho$ can be easily understood, taking into account that:

$$\sum_{l=0}^{\infty}\vec{q}_l\cdot\sum_{n=1}^{N}n\boldsymbol{D}_n\cdot\vec{e}_H^{\mathrm{T}}=\rho \tag{43}$$

Further on, having made the observations that:

$$\left[\sum_{i=n-l+1}^{N}\boldsymbol{D}_i\right]\cdot\vec{e}_H^{\mathrm{T}}=\sum_{i=n-l+1}^{N}\left[\boldsymbol{D}_i\cdot\left(\boldsymbol{G}^{i-n+l-1}\cdot\vec{e}_H^{\mathrm{T}}\right)\right] \tag{44}$$

and by identification with (25), we observe that:

$$\sum_{l=n-N+1}^{n}\vec{q}_l\cdot\left[\sum_{i=n-l+1}^{N}\boldsymbol{D}_i\right]\cdot\vec{e}_H^{\mathrm{T}}=q(n) \tag{45}$$

Similar observations can be made about the terms appearing in equations (40) and (41), so that finally we may state the significant result:

$$w(n)=\frac{1}{\rho}q(n)\qquad n\ge 1 \tag{46}$$

We have furthermore made the assumption in section 2.1 that when a batch of a_k cells arrives at time slot k, each cell has an equal and constant probability to be stored as the i-th cell within the batch into the queue, $1\le i\le a_k$. Consequently, the sojourn time PMF would be the same for a particular connection (the cells belonging to a particular connection are tagged).

2.8. Moments of the Sojourn Time Distribution

From expression (46), it is quite straightforward to obtain the sojourn time PGF, and its moments, as the PMFs $\tilde{Q}(z)$ and $\tilde{W}(z)$ are almost proportional:

$$\tilde{W}(z)=\frac{1}{\rho}\left[\tilde{Q}(z)-1+\rho\right] \tag{47}$$

and we also obtain directly the relationships between the i-th moment of the system contents and the waiting time:

$$W^{[i]} = \frac{Q^{[i]}}{\rho}, \quad i \geq 1 \tag{48}$$

which is nothing else than Little's law extended to all the moments of the sojourn time and system occupancy distributions! The relationship between the variance of the system occupancy and the variance of the sojourn time reads:

$$\mathrm{Var}[W] = \frac{\rho Q^{[2]} - \left(Q^{[1]}\right)^2}{\rho^2} = \frac{1}{\rho}\mathrm{Var}[Q] - \frac{1-\rho}{\rho^2}\left(Q^{[1]}\right)^2 \tag{49}$$

3. SUPERPOSITION OF ON-OFF SOURCES VIEWED AS A D-BMAP

We consider a superposition of a homogeneous set of discrete-time ON-OFF sources that have been introduced in [13] for the computation of the busy and idle periods, and have been used in various publications, e.g. [5] for the assessment of an ATM multiplexer performance under periodic input load.

3.1. The Discrete-Time ON/OFF Source Model

The discrete–time ON/OFF source model that we consider throughout this section requires three statistical parameters in order to be fully characterized. A natural formulation of the source characteristics resides in the triplet $\{\lambda, b, m\}$, where:

- λ is the per-source offered traffic;
- b is the average burst duration expressed in terms of the number of cells per burst;
- m is the minimum cell inter-arrival time, i.e. $r = 1/m$ is the source peak rate expressed relatively to the multiplexer output rate.

The OFF duration is multiple of one slot time and is geometrically distributed, with expectation $1/p$, and minimum duration one time slot. The ON phase consists of a geometrically distributed random multiple of m slots (minimum m time slots), in which the first one contains a single cell and the m–1 remaining slots are empty. The expectation of the number of "trains" of m consecutive slots is assigned the value $1/q$, and the ON state duration expectation becomes m/q slots. Consequently, p and q are the parameters of the geometric distributions involved in the state residence durations. The translation from parameter set $\{\lambda, b, m\}$ to the set $\{p, q, m\}$ reads as follows:

$$p = \frac{\lambda}{b(1-\lambda m)}, \quad q = \frac{1}{b} \tag{50}$$

This 3-parameter ON-OFF source can be modeled by an m+1 state MC where state 2 is the only one producing a cell when entered. State 1 represents the OFF state, while states 2 to m+1 model the ON state. States 3 to m+1 correspond to the m–1 empty slots following a generated cell in state 2. A corresponding D-BMAP formulation of this source model using matrices P_0

and P_1 can be obtained directly, where the transitions in column 2 are associated with P_1, and the other with P_0. The MPGF $P(z) = P_0+zP_1$ is given by:

$$P(z)=\underbrace{\begin{bmatrix} 1-p & pz & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \vdots & & & 1 \\ q & (1-q)z & 0 & \cdots & \cdots & 0 \end{bmatrix}}_{m+1} \tag{51}$$

and $P_0 = P(z{=}0)$, $P_1{=}P(z{=}1){-}P(z{=}0)$. When N such ON/OFF sources are superimposed, the total offered traffic becomes $\rho = N\lambda$. The parameters characterizing our queueing system is therefore upgraded to the set $\{p,q,m,N\}$. Two important loading cases need to be distinguished:

- *Underload* case: when the number of sources N is lower than or equal to the inverse of the peak rate, $1/r = m$, no multiplexing gain is achieved since the aggregate peak rate remains permanently lower or equal to the multiplexer output rate. We will refer to this case as the "underload" case hereafter. A direct consequence of the "underload" situation is that within a set of m consecutive slots, $N \leq m$ arrivals may occur, while at most m departures may occur. The number of cells in the system when $N \leq m$ may never override the upper bound of N cells in this case. The underload case is handled in section 3.3.
- *Overload* case: on the other hand, when $N > m$, the aggregate peak rate may temporarily override the one of the multiplexer output. This case will be referred to as an "overload" situation. Some multiplexing gain is achieved. The system occupancy may grow to any level larger than N.

We observe also, that in order to have some queueing, the number of superimposed sources needs to be $N \geq 2$.

3.2. Expression of an Aggregate of ON/OFF Sources as a D–BMAP

In this section, we formulate the expression of an aggregated set of N i.i.d. three-parameter ON/OFF sources. Then, from the previous results, the complete queueing system performance variables can be obtained. The D-BMAP is expressed using the transition probabilities between two overlapping arrival patterns. We first compute the state space size, i.e. the number of possible arrival patterns that can be obtained when superposing N ON/OFF sources of minimum cell inter-arrival time m. Then, we express the transition probability between two compatible arrival patterns.

3.2.1. Number of Arrival Patterns

The D–BMAP defined by D corresponding to the superposition of a set of N i.i.d. ON-OFF sources could be constructed as a simple combination of the individual ON-OFF source D-MAPs P [2] (for example, D_0 can be obtained by an N-fold Kronecker product of matrix P_0:

$D_0 = \otimes_{i=1}^{N} P_0$). However, the state space size reaching $(m+1)^N$ states would be absolutely prohibitive. Instead, we propose the usage of a sensibly reduced D–BMAP of size $H(N,m)$, where $H(N,m)$ corresponds to the number of possible arrival patterns $\vec{a}_i = \left[a_{i,1}\ \dots\ a_{i,m}\right]$, $1 \le i \le H(N,m)$, over m successive time slots, for a superposition of N i.i.d ON-OFF sources, where $H(N,m)$ can be computed as:

$$H(N,m) = \binom{N+m}{N} = \binom{N+m}{m} \tag{52}$$

In $\vec{a}_i = \left[a_{i,1}\ \dots\ a_{i,m}\right]$, $a_{i,j}$ denotes the number of cells in the j-th batch arrival within the interval of m consecutive slots, $1 \le j \le m$, in the i-th arrival pattern $\vec{a}_i$, according to a particular pattern ordering. Expression (52) corresponds to the typical problem of the number of configurations that can be obtained when putting a variable number of objects from 0 to N into m boxes. When the number of objects is known, i.e. say that there are i objects to be put into m boxes, the number of possible configurations $H_{i,m}$ can be computed using the following recursive algorithm:

$$H_{i,m} = \begin{cases} 1 & \text{for } m = 1 \\ \sum_{k=0}^{i} H_{k,m-1} & \text{for } m > 1 \end{cases} \tag{53}$$

and $H_{0,m} = H_{i,1} = 1$. Reducing the recursion in the previous expression provides the number of configurations available when putting i objects into m boxes, which yields:

$$H_{i,m} = \binom{i+m-1}{i} \tag{54}$$

However, in our case, a variable number of sources i between 0 and N inclusive may be active, so that $H(N,m)$ is the sum over i of $H_{i,m}$:

$$H(N,m) = \sum_{i=0}^{N} H_{i,m} = H_{N,m+1} = \binom{N+m}{N} \tag{55}$$

by virtue of the recursive relationship (53).

3.2.2. Transition Probabilities for Two Adjacent Patterns

Several contributions, e.g. [5], [13], exploited the fact that the transition probability from one arrival pattern $\vec{a}_i$ to the next one $\vec{a}_j$ for adjacent time slots can be easily formulated. We shall build a D-BMAP representing a QIP of the aggregate of ON/OFF sources. We define $\left[D_n\right]_{i,j}$ as the probability of obtaining pattern $\vec{a}_j$ at time slot k+1, knowing that pattern $\vec{a}_i$ was obtained at time slot k, and that n arrivals occur in the last slot of pattern $\vec{a}_j$ $(a_{j,m} = n)$. Let $s_i = \sum_{u=1}^{m} a_{i,u}$ be the sum of all arrivals within pattern $\vec{a}_i$; $0 \le s_i \le N$. The matrix $D = \sum_{n=0}^{N} D_n$ (as defined in (1)) of the D-BMAP of a superposition of a homogeneous set of

ON-OFF sources can be constructed as:

$$[D]_{i,j} = \left[D_{a_{j,m}}\right]_{i,j} = \begin{cases} T_{p,q,m,N}\left(a_{i,1}, s_i, a_{j,m}\right) & \text{when } a_{j,k-1} = a_{i,k}, \text{ for } 2 \le k \le m \\ & \text{and } 0 \le a_{j,m} \le N - s_i + a_{i,1} \\ 0 & \text{otherwise} \end{cases} \tag{56}$$

In our case, and for a given transition from a pattern $\vec{a}_i$ to another pattern $\vec{a}_j$, a unique number of cells can be produced, since an arrival pattern finishes with a unique number of arrivals $a_{j,m}$. This observation together with expression (56) also implies that elements (i, j) in the matrices $\boldsymbol{D}_n$, $0 \le n \le N$, whose associated patterns do not finish with n arrivals equal forcibly zero:

$$\left[D_n\right]_{i,j} = 0, \quad \text{when } n \ne a_{j,m} \tag{57}$$

Also, the destination pattern of the transition must include the history of arrivals of the previous pattern, and only the last (new) number of arrivals may be variable, i.e. arrivals $a_{i,2}$ to $a_{i,m}$ must match arrivals $a_{j,1}$ to $a_{j,m-1}$. The transition probabilities $T_{p,q,m,N}$ $(a_{i,1}, s_i, a_{j,m})$ result from the convolution of the distribution of two contributions. The next number of arrivals $a_{j,m}$ equals the sum k_1+k_2 of the number of sources k_1, $0 \le k_1 \le a_{i,1}$, that were active m time slots before and remain active (each such event occurs with probability 1–q):

$$\mathrm{P}\left[k_1 \text{ sources remain active}\right] = \binom{a_{i,1}}{k_1}(1-q)^{k_1} q^{a_{i,1}-k_1}, \quad 0 \le k_1 \le a_{i,1} \tag{58}$$

and of the number of sources k_2, $0 \le k_2 \le N$–s_i, that were not active one time slot before and that become active (each such event happens with probability p):

$$\mathrm{P}\left[k_2 \text{ sources become active}\right] = \binom{N-s_i}{k_2} p^{k_2}(1-p)^{N-s_i-k_2}, \quad 0 \le k_2 \le N - s_\mathrm{i} \tag{59}$$

since each active source produces exactly one cell in a window of m successive ATM slots (due to the enforced spacing of m–1 empty slots between cells), so that the number of cells in any such window gives the number of sources that are active at the end of this window period. The probability that $a_{j,m} = k_1$+k_2 sources produce a cell in the slot following the last slot of the observation window is then the convolution of PMFs (58) and (59), since all sources are considered independent:

$$T_{p,q,m,N}\left(a_{i,1}, s_i, a_{j,m}\right) = \sum_{k=\mathrm{Max}\left[0, a_{j,m}-N+s_i\right]}^{\mathrm{Min}\left[a_{j,m}, a_{i,1}\right]} \binom{a_{i,1}}{k}\binom{N-s_i}{a_{j,m}-k}(1-q)^k q^{\left(a_{i,1}-k\right)} p^{a_{j,m}-k}(1-p)^{N-s_i-a_{j,m}+k} \tag{60}$$

The transition probabilities depend only on $a_{i,1}$, s_i and $a_{j,m}$ as can be seen in (60).

3.2.3. Number of Boundary Probabilities and Sparsity of the Matrices Involved

Considering the conditions in (56) for a transition probability to be non zero, one sees that $\boldsymbol{D}$ becomes more and more sparse with increasing m. When m=1, the conditions in (56) are always true and the $(N+1)^2$ transition probabilities in $\boldsymbol{D}$ are all non zero.

The B–DMAP formulation of the problem based on the source arrival patterns yields some immediate results. First, the steady state of the underlying source process MC of matrix $\boldsymbol{D}$ can be directly expressed by a multinomial PMF. The steady state probability of state i equals the probability of apparition of an arrival pattern $\vec{a}_i$:

$$\left[\vec{d}\right]_i = \mathrm{P}\left[\vec{a}_i\right] = \frac{N!}{\left(N-s_i\right)!\prod_{n=1}^{m} a_{i,n}!}\lambda^{s_i}\left(1-m\lambda\right)^{N-s_i} \tag{61}$$

Second, all arrival patterns are not compatible with any queue size l. In particular, the patterns $\vec{a}_i$ that allow an empty system must, among other conditions, have their last batch arrival $a_{i,m}$ equal to zero according to (2), since there remains at least $a_{i,m}$ cells in the system when the m-th batch arrival pertaining to pattern i occurs. In fact, the function giving the amount of different arrival patterns that allow queue size l and possibly higher in the overload case (but not lower), can be described by a function $F_l(N,m)$ of the number of sources and the minimum cell inter-arrival time m. This function is analyzed in the following.

Clever observations of the pattern generation process lead to the determination of the function F for overload and underload cases. This function plays an important role, since it will provide us with a compact algorithm for the computation of the system occupancy distribution in the "underload" cases (see section 3.3), while F=$F_0(N,m)$ informs us directly about the number of boundary probabilities in a particular queueing system (see end of this section) that need to be solved in order to derive all performance measures. We illustrate hereafter the determination of $F_0(N,m)$ in the "overload" case; other values of $F_l(N,m)$ can be derived similarly, at the expense of tedious reflections.

Let us monitor our multiplexer system during a window of m contiguous time slots such that the system is empty after the batch of arrivals occurring during the last slot in the window, and tag these slots from slot number k+1 up to slot number k+m. As we are in an "overload" case, we have $N > m$, and one expects that the number of patterns allowing an empty system should not depend on N, since a maximum of k+m–j arrivals may appear at time slot j (for $k+1 \leq j \leq k+m$), i.e. j–1 time slots before the system is required to be empty (time slot k+m), in order to flush completely the arrivals (otherwise the system may not return to the empty state even if it was empty at time slot k). The number of allowable arrivals depends then on the time left until "system empty" status and not on the amount of sources N, and let us call this particular case of the function $F_0(N,m)$ as $F_0(\infty,m)$. The last number of arrivals occurring at time slot k+m must therefore always be zero. The patterns allowing an empty system all terminate with zero arrivals, and clearly $F_0(\infty,1)$=1, i.e. $\vec{a}_1 = [0]$ is the only arrival "pattern" compatible with an empty system in the case m=1.

For m>1, patterns allowing an empty system and corresponding to a spacing of m+1 can be constructed recursively, starting from those for a spacing of m, as illustrated on figure 3. The number of patterns at "depth" m results from a series of embedded summations as described by the expression:

$$F_0(\infty,m) = \sum_{j_1=1}^{1}\sum_{j_2=2}^{j_1+1}\sum_{j_3=2}^{j_2+1}\cdots\sum_{j_m=2}^{j_{m-1}+1} j_m \tag{62}$$

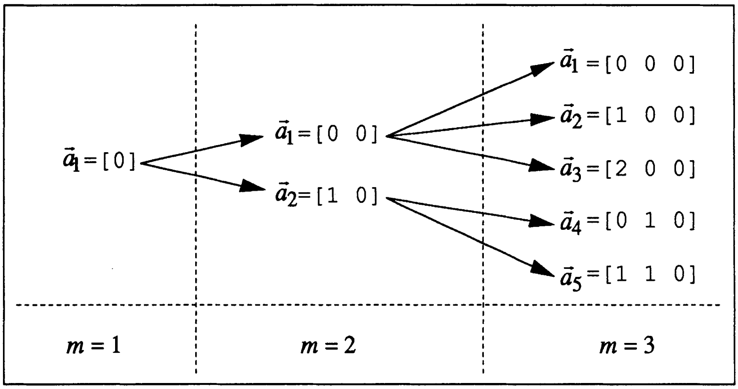

Figure 3 : illustration of the recursive construction of arrival patterns allowing an empty system up to m=3.

Reducing level by level the summations in (62) reveal the final function $F_0(\infty,m)$ as can be seen in (63) and (64). Similar observations lead to the general function $F_l(N,m)$ that is presented hereafter:

$$F_l(N,m)=\begin{cases} N>m \begin{cases} 0\le l\le N-m, & \binom{2m+l}{m+l+1}\dfrac{l+1}{m} \\ N-m\le l\le N-1, & H(N,m)-\binom{N+m}{N-l-1} \\ N\le l, & H(N,m) \end{cases} \\ N\le m \begin{cases} 0\le l\le N-1, & \binom{N+m}{N-l}-\binom{N+m}{N-l-1} \\ l=N & 1 \\ l>N & 0 \end{cases} \end{cases} \tag{63}$$

In (63), overlapping conditions yield the same value. When observing the number of arrival patterns that allow an empty system only, a more handsome formulation can be obtained, i.e.:

$$F=F_0(N,m)=\begin{cases} \binom{N+m}{N}\dfrac{m-N+1}{m+1}, & \text{for } N\le m-1 \\ \dfrac{1}{m}\binom{2m}{m+1}, & \text{for } N\ge m-1 \end{cases} \tag{64}$$

From now on, it is assumed that the patterns are sorted within the D-BMAP according to an increasing allowance l of the number of cells in the system, i.e. patterns from 1 to $F_0(N,m)$ allow an empty system and any other queue size in the "overload" case; patterns $F_0(N,m)+1$ to $F_1(N,m)$ allow queue sizes 1 and higher, and so on up to pattern number $H(N,m)$ that allows queue size(s) N (and higher in the "overload" case): $\vec{a}_{H(N,m)} = [0 \;\cdots\; 0 \;\; N]$.

As a direct consequence of these observations, the number of non-zero boundary value elements in $\vec{q}_0$ match exactly $F = F_0(N,m)$. Thank to our arrival pattern ordering, we deduce that $[\vec{q}_0]_j = 0$ for $j > F$. Moreover, a "first passage" can only occur with a pattern $\vec{a}_j$, $1 \le j \le J(N,m)$, finishing with zero arrivals. Consequently, the "first passage" matrix G encompasses exactly $H(N,m)$–$J(N,m)$ zero columns, where:

$$J(N,m) = H(N,m-1) \tag{65}$$

is the number of arrival patterns that finish with zero arrivals (the F arrival patterns that allow an empty system are a subset of the $J(N,m)$ patterns that terminate with zero arrivals). We have observed as well that most of the non-zero elements of G are grouped within its F first columns. The steady state vector of G needs therefore to be computed only from the submatrix G_F composed by the first F rows and columns of G. Let $\vec{g}_F$ be this vector defined by $\vec{g}_F \cdot G_F = \vec{g}_F$ and $\vec{g}_F \cdot \vec{e}_F^{\,T} = 1$; then clearly $\vec{g} = [\vec{g}_F \;\; \vec{0}]$.

Also, the contents of the $G(0)$ matrix needs to be properly initialised in order to ensure that the intermediate matrices $G(k)$ in the computation of (20) remain *sparse* ones. We refer to [14], for a complete algorithm for the determination of the non-zero elements in G.

3.3. Solution of the Queue Distribution in the "underload" Case

We have defined earlier an "underload" condition of the queueing system corresponding to the case where $N \le m$, i.e. the aggregate peak rate of the sources cannot exceed the output rate of the multiplexer and therefore no multiplexing gain is achieved. In this case, the maximum number of cells in the system cannot exceed N, since up to m departures may occur within m slots, and the corresponding number of arrivals is lower or equal to that value.

When $N < m$, and for the same reasons, within each period of m consecutive slots, the system returns to empty at least once. The system contents at the end of each pattern depends therefore only on the arrivals since the point in time where the system was empty, and not on the size at the beginning of the pattern. When $N = m$, the systems returns at least once during every period of m slots to a level equal or less to "once cell in system". Consequently it is sufficient to know the history of arrivals during the last m slots in order to determine the system contents at the end of the period, i.e. each arrival pattern $\vec{a}_i$ is associated with just one system size obtained by applying recursively (2) m times, starting with n_0=0 and mapping each a_k with $a_{i,k}$. Let us define the additional function $K_l(N,m)$, giving the number of arrival patterns that are compatible with system contents up to level l:

$$K_l(N,m) = \begin{cases} 0 & l = -1 \\ \sum_{j=0}^{l} F_j(N,m) & 0 \le l \le N \end{cases} \qquad -1 \le l \le N \tag{66}$$

The handy ordering of the arrival patterns we designed previously enables us to write the probability of a number l of cells in the system q_l in the following compact form :

$$q(l) = \sum_{j=K_{l-1}(N,m)+1}^{K_l(N,m)} \mathrm{P}[\vec{a}_i] \tag{67}$$

and the steady state vector elements of the submatrix G_F are given also directly:

$$[\vec{g}_F]_j = \frac{1}{1-\rho} \mathrm{P}[\vec{a}_j], \quad 1 \le j \le F \tag{68}$$

Consequently, in the "underload" case, $\sum_{i=1}^{F} \mathrm{P}[\vec{a}_i] = 1-\rho$.

4. SUBCASES OF THE ON/OFF SOURCE REVISITED

Based on the results of the previous section about ON/OFF sources, we can derive some simplified expressions for the following parameter restrictions of the ON/OFF source model:

- the 2 State Markov source (2SM) that corresponds to the restriction m = 1 in the general ON/OFF source model and;
- the Bernoulli source model further restricted with the condition p+q=1.

The handling of queueing process involving 2SM and Bernoulli sources drastically differs from the problematic of the general ON/OFF sources superposition because in the former 2 cases, the boundary probability is unique and can be determined beforehand. Solving the boundary probabilities using (20) is not necessary, therefore most of the results can be expressed in closed-form expressions. We shall derive in the next two subsections striking results for these two subcases.

4.1. Simplifications for the Subcase of a Superposition of 2SM Sources

When the parameter restriction m = 1 is applied to our ON-OFF source model, we obtain the so-called 2SM source model, whose peak rate equals the multiplexer output rate,. There are exactly N+1 possible arrival patterns (check with (52)), i.e. from pattern $\vec{a}_1 = [0]$ to pattern $\vec{a}_{N+1} = [N]$, among which only the first one $\vec{a}_1 = [0]$ is compatible with an empty system (check with (64)). Matrix $\boldsymbol{D}$ is therefore of dimension (N+1)x(N+1), whose elements are all non zero under normal circumstances. Consequently, we have $[\vec{g}]_1 = g_1 = 1$, $g_i = 0$, $2 \le i \le N+1$, $\mathrm{P}[Q=0, X=1] = [\vec{q}_0]_1 = 1-\rho$, and $[\vec{q}_0]_j = 0$ for $2 \le j \le N+1$ in this case.

As we consider arrival patterns over m = 1 slots only, s_i reduces to $s_i = a_{i,1}$ in equation (60). The elements in the matrix $\boldsymbol{D}$ can be immediately expressed as:

$$[\boldsymbol{D}]_{i+1,j+1} = \sum_{k=\mathrm{Max}[0,i+j-N]}^{\mathrm{Min}[i,j]} \binom{i}{k}\binom{N-i}{j-k}(1-q)^k q^{(i-k)} p^{j-k} (1-p)^{N-i-j+k}, \quad 0 \le i,j \le N \tag{69}$$

Of course, as previously, the j-th column of $\boldsymbol{D}$, $1 \le j \le N+1$, is the same as the j-th column in matrix $\boldsymbol{D}_{j-1}$, since arrival pattern $\vec{a}_j = [j-1]$ generates j–1 arrivals. The j-th column in the remaining N matrices $\boldsymbol{D}_n$, $0 \le n \le N$, $n \ne j-1$, equals the (N+1)x1 column vector of zeroes $\vec{0}_{N+1}^{\mathrm{T}} = [0 \ \cdots \ 0]^{\mathrm{T}}$. Element i in vector $\vec{d}$ follows a simple binomial distribution $b_1(N,\lambda;i\text{-}1)$:

$$b_1(N,p;k)=\begin{cases}0 & \text{for } k<0 \text{ or } k>N \\ \binom{N}{k}p^k(1-p)^{N-k} & \text{for } 0\le k\le N\end{cases} \tag{70}$$

of parameters the number of superimposed sources N and the per-source offered traffic λ:

$$\left[\vec{d}\right]_i=\binom{N}{i-1}\lambda^{i-1}(1-\lambda)^{N-i+1}, \qquad 1\le i\le N+1 \tag{71}$$

where the per-source offered traffic λ (sec. 3.1) reduces to $\lambda = p/(p+q)$.

The only arrival pattern allowing an empty system is also the only one that is the target of a "first passage" (zero number arrivals). Therefore, the $\boldsymbol{G}$ matrix can be immediately characterised as follows : $[G]_{i,1} = 1$, and $[G]_{i,j} = 0$, for $2 \le j \le N+1$ (some queueing occurs only when $N \ge 2$), and $1 \le i \le N+1$. A direct consequence is that $\boldsymbol{G}^n = \boldsymbol{G}$, $n \ge 1$, when $m = 1$. The computation of the matrices $\boldsymbol{V}(n)$ defined in expression (26) can be simplified as follows:

$$\boldsymbol{V}^*(n)=\sum_{i=n}^{N}\boldsymbol{D}_i\cdot\boldsymbol{G}^{i-n}=\begin{cases}\boldsymbol{D}_n+\left(\sum_{j=n+1}^{N}\boldsymbol{D}_j\right)\cdot\boldsymbol{G} & \text{when } 1\le n<N \\ \boldsymbol{D}_N & \text{when } n=N\end{cases} \tag{72}$$

A special form can be obtained for $\boldsymbol{V}^*(1)$:

$$\boldsymbol{V}^*(1)=\left(\boldsymbol{I}_{N+1}-\boldsymbol{D}_1\right)\cdot\boldsymbol{G}+\boldsymbol{D}_1-\boldsymbol{D}_0 \tag{73}$$

The system contents can be easily computed from (27)-(29). The first two moment can also be developed easily and a closed form formula expressed, thank to the unique boundary probability in the 2SM case. A handsome formulation for the first moment of the queue occupancy distribution has already been established [9] (this result has been obtained by other authors in different forms, e.g. [15], [4], [11]):

$$Q^{[1]}=\rho+\frac{\rho^2}{2(1-\rho)}G(p,q)\left(1-\frac{1}{N}\right) \tag{74}$$

where G is the source granularity [9]:

$$G(p,q)=\frac{2}{p+q}-1=2b(1-\lambda)-1 \tag{75}$$

where $\{b, \lambda\}$ and $\{p, q\}$ have been related to each other in expression (50). A closed form expression can be found also for the variance of the queue occupancy:

$$\begin{aligned}\mathrm{Var}[Q] &= \rho(1-\rho)+\frac{\rho^2}{12(1-\rho)^2}\frac{N-1}{N^2}\Big\{3G(p,q)^2(N-1)+\\ &+(1-\rho)\Big[(1-\rho)\big(3G(p,q)^2+12G(p,q)+1\big)+2-\\ &-2(N-1)\big(1-3G(p,q)^2\big)-(1-\rho)(N-1)\big(6G(p,q)^2-12G(p,q)-5\big)\Big]\Big\}\end{aligned} \tag{76}$$

Using relationships (48), we obtain:

$$W^{[1]}=1+\frac{\rho}{2(1-\rho)}G(p,q)\left(1-\frac{1}{N}\right) \tag{77}$$

$$\begin{aligned}\mathrm{Var}[W] &= \frac{\rho}{4(1-\rho)^2}G(p,q)^2\left(1-\frac{1}{N}\right)^2-\\ &-\frac{\rho}{12(1-\rho)}\frac{N-1}{N^2}\Big\{\rho\big[3G(p,q)^2(2-N)+5N-4\big]-3\big(G(p,q)^2+N\big)\Big\}\end{aligned} \tag{78}$$

4.2. Simplifications for the Subcase of a Superposition of Bernoulli sources

With the further restriction $p+q=1$, while keeping the restriction $m = 1$, a superposition of N Bernoulli sources is obtained. Successive batch arrivals are uncorrelated so that simple operations on PGFs lead to the PGF of the system occupancy:

$$\tilde{Q}(z)=\frac{(1-\rho)(z-1)\tilde{A}(z)}{z-\tilde{A}(z)} \tag{79}$$

where $\tilde{A}(z)$ corresponds to the PGF of the number of arrivals at an arbitrary slot, which is a binomial PGF for a superposition of N Bernoulli sources:

$$\tilde{A}(z)=(1-p+pz)^N \tag{80}$$

We may also apply the results in section 2.7 to the case of the superposition of Bernoulli sources. These operations yield the sojourn time PGF (compare with (47)):

$$\tilde{W}(z)=\frac{1}{\rho}\left[\tilde{Q}(z)-(1-\rho)\right]=\frac{1}{\rho}\frac{(1-\rho)z\left[\tilde{A}(z)-1\right]}{z-\tilde{A}(z)} \tag{81}$$

and:

$$Q^{[1]}=\rho+\frac{\rho^2}{2(1-\rho)}\left(1-\frac{1}{N}\right) \tag{82}$$

$$\mathrm{Var}[Q] = \rho(1-\rho) + \frac{\rho^2}{12(1-\rho)^2}\frac{N-1}{N^2}\left(18N - 8\rho - 26N\rho + 5\rho^2 + 11N\rho^2\right) \tag{83}$$

and the moments of the sojourn time equal :

$$W^{[1]} = \frac{Q^{[1]}}{\rho} = 1 + \frac{\rho}{2(1-\rho)}\left(1 - \frac{1}{N}\right) \tag{84}$$

$$\mathrm{Var}[W] = \frac{\rho}{12(1-\rho)^2}\frac{N-1}{N^2}\left[6N + \rho(2\rho - 5)(N+1)\right] \tag{85}$$

5. NUMERICAL EXAMPLES

For illustration purposes, we have resolved a couple of "small" cases corresponding to $\lambda = 0.08$, $N = 10$, $m = 5$ and variable mean burst sizes b (A cases in figure 7). The typical "knee" that can be observed in curves representing the loss probability versus buffer size for limited-size multiplexers (e.g. in [5]) appears in the queue length PMF as well. It appears mainly for higher mean burst sizes, i.e. 10, 20 and 50 (figure 4) in our case. The "knee" delimits two regions corresponding to the "cell scale" congestion for small queue occupancies (less than m sources active) and the "burst scale" congestion for large queue occupancies (more than m sources active). As it is already well known, the decay after the "knee" reduces with the increase of b.

The cell sojourn time complementary cumulative distribution function for a couple of 2SM source loadings has been illustrated on graph 5 (B cases in figure 7). The moments of the contents and the sojourn time for the A cases can also be obtained using expressions (37), (39), and (48) or directly from the boundary values and the source parameters using the results in [12] (see graph 6).

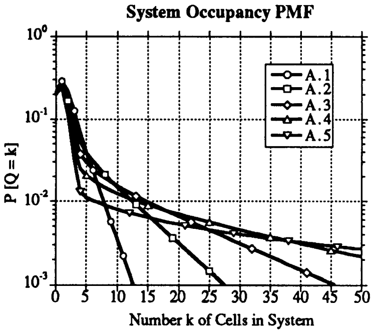

Figure 4: System contents PMF, P[Q=k] for an overall offered traffic ρ=0.8, with N=10 sources and m=5. The mean burst size b is varied between 2 and 50 cells/burst.

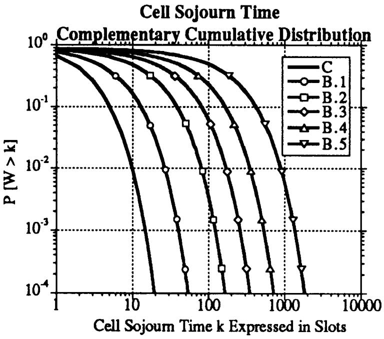

Figure 5: Similar case to graph 4. Sojourn time complementary cumulative distribution function for 2SM sources, an overall offered traffic ρ=0.8, with N=10 sources. The mean burst size b is also varied between 2 and 50 cells/burst.

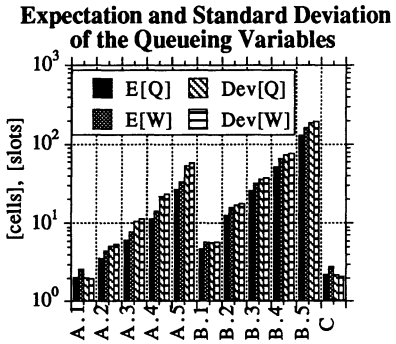

Figure 6: Expectation and standard deviation of the number of cells in the system Q, and of the cell sojourn time W.

Case	ρ	N	m	b
A.1	0.8	10	5	2
A.2	0.8	10	5	5
A.3	0.8	10	5	10
A.4	0.8	10	5	20
A.5	0.8	10	5	50
B.1	0.8	10	1	2
B.2	0.8	10	1	5
B.3	0.8	10	1	10
B.4	0.8	10	1	20
B.5	0.8	10	1	50
C	0.8	10	1	25/23

Figure 7: Table describing the testcase parameters. Cases A.1 to A.5 correspond to a superposition of homogeneous ON/OFF sources, B.1 to B.5 to 2SM sources, and C to Bernoulli sources.

The number of arrival patterns $\vec{a}_1$ to $\vec{a}_{H(N,m)}$ patterns corresponding to the A cases computed above reach $H(10,5) = 3'003$. The number of boundary patterns equal $F(10,5) = 42$. The number of non zero elements in the $\boldsymbol{G}$ matrix equals 118'954 over a total amount of 3'003x3'003 elements, which yields a density of 1.32%.

6. CONCLUSIONS

Correlated source models representing the ATM input traffic have been developed and so was the suitable D–BMAP source model. We have summarized the expressions that solve completely the queue distribution (equations (27) to (30)) and the first two moments (equations (37) to (39)) of a general D–BMAP/D/1 queueing system, that is often used as a basis for the computation of ATM multiplexers performance. We have then established the relationships between system contents distribution and sojourn time distributions, as well as the moments ((40) to (49)).

Based on these first results, we have defined the D–BMAP that models a superposition of a small set of three-parameter ON/OFF sources, explained the sparsity of the matrices and vectors involved while computing the boundary probabilities. Subcases of the ON/OFF sources, such as the underload cases (sect. 3.3), where no multiplexing gain is achieved, or a homogeneous superposition of two-state Markov sources and Bernoulli sources have been fully handled (sect. 4).

The contribution of this work is threefold; first we have extensively derived from previous results the solution of a D-BMAP/D/1 queue (queue length and sojourn time distributions), with service time unity. These results are particularly relevant in the ATM context. Second, we have obtained exact results (to the extent of an exact computation of (20)) for the queue distribution when a small set of high bitrate ON/OFF sources load a switching/multiplexing element. These results might be particularly relevant for equipment testing purposes. Finally, using the results on the moments together with results provided in [12], we are able to express directly from the boundary probabilities the delay and jitter of a group of homogeneous connections. Another significant result resides in the observations made about the "underload" cases: the computation of the boundary probabilities can be bypassed in these cases, and the moments obtained immediately.

ACKNOWLEDGEMENTS

The author would like to thank Prof. P.–G. Fontolliet for carefully reviewing this paper. Many thanks also to Prof. Dr. Chris Blondia, University of Nijmegen, and Uwe Briem, University of Stuttgart, for very fruitful discussions during RACE project meetings.

REFERENCES

[1] A. Baiocchi et al., "Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed On-Off Sources", *IEEE JSAC*, vol. 9, no. 3, pp. 388-393, April 1991.

[2] C. Blondia and O. Casals, "Statistical Multiplexing of VBR Sources: a Matrix-Analytic Approach", *Performance Evaluation*, vol. 16, pp. 5-20, 1992.

[3] U. Briem, T. Theimer, and H. Kröner, "A General Discrete-Time Queueing Model: Analysis and Applications", in *International Teletraffic Congress 13 (ITC-13)*, pp. 13-19, A. Jensen and V.B.I. editor , Elsevier Science Publishers B.V., 1991.

[4] H. Bruneel, "Queueing Behavior of Statistical Multiplexers with Correlated Inputs", *IEEE Transactions on Communications*, vol. 36, no. 12, pp. 1339-341, December 1988.

[5] H. Kröner, "Statistical Multiplexing of Sporadic Sources - Exact and Approximate Performance Analysis", in *Proc. of 13th International Teletraffic Congress (ITC 13)*, pp. 787-793, A. Jensen and V.B.I. editor , Elsevier Science Publishers B.V., Copenhagen, Denmark, June 1991.

[6] S.Q. Li, "Generating Function Approach for Discrete Queueing Analysis with Decomposable Arrival and Service Markov Chains", in *INFOCOM'92*, pp. 2168-2177, Florence, Italy, May 6-8 1992.

[7] D.M. Lucantoni, "New Results on the Single Server Queue with a Batch Markovian Arrival Process", *Communications in Statistics - Stochastic Models*, vol. 7, pp. 1-46, 1991.

[8] D.M. Lucantoni, *BMAP/G/1 Queue: A Tutorial*, vol. 729, Lecture Notes in Computer Science. L. Donatiello and R. Nelson Ed., pp. 330-358.

[9] M. Luoni, "ATM Traffic Characterization with Applications to Connection Acceptance Control", Ph.D. thesis, Thèse no 979 (1991), Ecole Polytechnique Féd. de Lausanne, DE-TCOM, 1991.

[10] M.F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and their Applications*. New York and Basel: Marcel Dekker, Inc., 1989, Probability: Pure and Applied - A series of Textbooks and Reference Books.

[11] M. Neuts, "On Viterbi's Formula for the Mean Delay in a Queue of Data Packets", *Commun.Statist.-Stochastic Models*, vol. 6, no. 1, pp. 87-98, 1990.

[12] R. Slosiar, "Moments of the Queue Occupancy in an ATM Multiplexer Loaded with ON/OFF Sources", in *Proc. of the IEEE Singapore International Conference on Communication Systems (ICCS '94)*, vol. 2/3, pp. 754-759, November 14-18 1994.

[13] R. Slosiar, “Busy and Idle Periods at an ATM Multiplexer Output Resulting from the Superposition of Homogeneous ON/OFF Sources", in *Proc. of 14th Internantional Teletraffic Congress*, vol. vol. 1a, pp. 431-440, J. Labetoulle and J.W.R. Editors, Elsevier publ., Antibes, France, 6-10th June 1994.

[14] R. Slosiar, “Performance Analysis Methods of ATM-Based Broadband Access Networks using Stochastic Traffic Models", Ph.D. thesis, to appear, EPFL, DE-TCOM, CH-1015 Lausanne, 1995.

[15] A.M. Viterbi, “Approximate Analysis of Time-Synchronous Packet Networks", *IEEE JSAC*, vol. 4, no. 6, pp. 879-890, April 1986.

[16] Y. Xiong, “Analysis of the Asymptotic Queueing Behavior of Statistical Multiplexers with General Markov-Modulated Traffic Sources", Ph.D. thesis, Faculty of Applied Sciences, University of Gent, 1994.

[17] Z. Zhang, “Analysis of a Discrete-Time Queue with Integrated Bursty Inputs in ATM Networks", *International Journal of Digital and Analog Communication Systems*, vol. 4, pp. 191-203, 1991.