On the Queue Length Distribution in BMAP Systems

Andrzej Chydzinski

***Abstract*— Batch Markovian Arrival Process – BMAP – is a teletraffic model which combines high ability to imitate complex statistical behaviour of network traces with relative simplicity in analysis and simulation. It is also a generalization of a wide class of Markovian processes, a class which in particular include the Poisson process, the compound Poisson process, the Markov-modulated Poisson process, the phase-type renewal process and others. In this paper we study the main queueing performance characteristic of a finite-buffer queue fed by the BMAP, namely the queue length distribution. In particular, we show a formula for the Laplace transform of the queue length distribution. The main benefit of this formula is that it may be used to obtain both transient and stationary characteristics. To demonstrate this, several numerical results are presented.**

***Index Terms*— teletraffic modeling, finite-buffer queue, BMAP, performance evaluation**

I. Introduction

The batch Markovian arrival process was invented by Neuts and in the beginning it was called the versatile Markovian point process or the N-process. Then, Lucantoni replaced the original complicated parameterization of the N-process by a new one, more simple and intuitive. Since then the process has been called the batch Markovian arrival process (BMAP).

The BMAP is a tool of choice for traffic modeling and predictability, performance evaluation of buffering processes, congestion and admission control mechanisms etc.

It is worth recommending for several reasons. Firstly, it is able to mimic the self-similar and bursty nature ([3], [4]) of network traces, remaining analytically tractable due to its Markovian structure.

Secondly, the BMAP generalizes a wide set of processes used in teletraffic modeling. For instance, by setting the proper parameterization of the BMAP we may obtain a Poisson process, a batch Poisson process, a Markov-modulated Poisson Process (MMPP), a Markovian arrival process (MAP) or a phase-type renewal process. Some of them are classic, others, like MMPP, gained great attention in applications connected with multimedia and ATM [5]–[9]. All results obtained for the BMAP can be automatically used for the processes mentioned above.

Thirdly, the BMAP is successfully used for the modeling of aggregated IP traffic [10], [11]. In this approach, different lengths of IP packets are represented by BMAP batch sizes. What is important is that algorithms for the fitting of the BMAP parameters to recorded IP traces are available [10], [11]. Some examples of particular performance issues connected with IP networks, analyzed by means of the BMAP, can be found by the reader in [12], [13].

In this paper we deal with the basic characteristic of a single-server queueing system, namely the queue size distribution. The queue size distribution and its parameters (average, variance) play an important role in the performance evaluation of buffering mechanisms in network devices by providing fundamental insight into the system's behaviour.

The main contribution of this paper is the Laplace transform of the transient queue size distribution in the BMAP queue with finite buffer (Theorem 1). To the best of the author's knowledge, there have been no reported results of this type yet. The classic papers by Ramaswami and Lucantoni [2], [14]–[16] are devoted to BMAP queues with infinite buffers. Articles in which the finite-buffer model is investigated are rare and deal with stationary characteristics [17] or special cases of the BMAP, like MMPP [18], only.

The original approach applied in this paper gives the results in a closed form which permits one to easily compute the stationary as well as the time-dependent queue size distribution. All the analytical results obtained herein were checked and confirmed by means of a discrete-event simulator written in OMNET++ [19].

The paper is organized as follows. Firstly, the model of the queue and the arrival process are presented and the notation is listed (section II). Then, in section III, the formula for the transform of the transient queue size is proven. Furthermore, some remarks on how it can be used in practice for obtaining time-dependent and stationary queue size distributions are given. In section IV, a set of numerical examples based on four different BMAP parameterizations is shown. In addition, computational aspects connected with coefficient matrices occurring in the main formula are discussed. Finally, conclusions are gathered in section V.

II. Queueing model and notation

In the paper we investigate a single server queueing system whose arrival process is given by a BMAP. The service time is distributed according to a distribution function $F(\cdot)$, the buffer size (queueing capacity) is finite and equal to b (including service position). In Kendall's notation, the system described is denoted by $BMAP/G/1/b$.

As regards the BMAP, it is constructed by considering a 2-dimensional Markov process $(N(t), J(t))$ on the state space $\{(i, j) : i \geq 0, 1 \leq j \leq m\}$ with an infinitesimal generator Q

Manuscript received July 07, 2006. This is an extended version of the paper presented during NEW2AN conference, Sankt Petersburg, May 2006.

A. Chydzinski is with the Silesian University of Technology, Institute of Computer Sciences. Address: Akademicka 16, 44-100 Gliwice, Poland. Email: Andrzej.Chydzinski@polsl.pl

1845-6421/06/6140 © 2006 CCIS

in the form:

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 & \cdot\cdot \\ & D_0 & D_1 & D_2 & \cdot\cdot \\ & & D_0 & D_1 & \cdot\cdot \\ & & & \cdot & \cdot\cdot \end{bmatrix},$$

where D_k, $k \geq 0$ are $m \times m$ matrices. D_k, $k \geq 1$ are nonnegative, D_0 has nonnegative off-diagonal elements and negative diagonal elements and

$$D = \sum_{k=0}^{\infty} D_k$$

is an irreducible infinitesimal generator (see [2]). It is assumed that $D \neq D_0$. Variate $N(t)$ represents the total number of arrivals in $(0,t)$, while variate $J(t)$ represents the auxiliary state (phase) of the modulating Markov process.

If π denotes the stationary vector for D ($\pi D = \mathbf{0}$, $\pi\mathbf{1} = 1$) and **1** stands for the column vector of 1's, then the total average intensity of the BMAP can be calculated as:

$$\Lambda = \pi \sum_{k=1}^{\infty} k D_k \mathbf{1}, \tag{1}$$

while the intensity of group arrivals as:

$$\Lambda_g = \pi(-D_0)\mathbf{1}. \tag{2}$$

The variance of the interarrival time is given by:

$$Var = -\frac{2}{\Lambda_g}\pi D_0^{-1}\mathbf{1} - \frac{1}{\Lambda_g^2}. \tag{3}$$

Finally, the correlation between two consecutive interarrival times can be calculated as follows:

$$\rho = \frac{1}{Var}\left(\frac{1}{\Lambda_g}\pi D_0^{-1}(D - D_0)D_0^{-1}\mathbf{1} - \frac{1}{\Lambda_g^2}\right). \tag{4}$$

An alternative, constructive definition of a BMAP is the following. Assume the modulating Markov process is in some state i, $1 \leq i \leq m$. The sojourn time in that state has exponential distribution with parameter λ_i. At the end of that time there occurs a transition to another state and/or the arrival of a batch. Namely, with probability $p_i(j,k)$, $1 \leq k \leq m$, $j \geq 0$, there will be a transition to state k with a batch arrival of size j. It is assumed that:

$$p_i(0,i) = 0,$$

and

$$\sum_{j=0}^{\infty}\sum_{k=1}^{m} p_i(j,k) = 1, \qquad 0 \leq i \leq m.$$

The relations between parameters D_k and λ_i, $p_i(j,k)$ are the following:

$$\begin{aligned} \lambda_i &= -(D_0)_{ii}, && 1 \leq i \leq m, \\ p_i(0,k) &= \frac{1}{\lambda_i}(D_0)_{ik}, && 1 \leq i,k \leq m, \quad k \neq i, \\ p_i(j,k) &= \frac{1}{\lambda_i}(D_j)_{ik}, && 1 \leq i,k \leq m, \quad j \geq 1. \end{aligned}$$

In the sequel, the following notation will be of use:

$\mathbf{P}(\cdot)$ – the probability

$X(t)$ – the queue size at the moment t (including service position if not empty)

$P_{i,j}(n,t) = \mathbf{P}(N(t){=}n, J(t){=}j \mid N(0){=}0, J(0){=}i)$ – the counting function for the BMAP. $N(t)$ denotes the total number of arrivals in $(0,t)$

$a_{k,i,j}(s){=}\int_0^{\infty} e^{-st} P_{i,j}(k,t)dF(t)$,

$f(s) = \int_0^{\infty} e^{-st}dF(t)$ – the transform of the service time distribution

δ_{ij} – the Kronecker symbol ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise)

In addition, we will be using the following $m \times m$ matrices:

$$\begin{aligned}
I &= m \times m \text{ identity matrix}, \\
\mathbf{0} &= m \times m \text{ matrix of zeroes}, \\
A_k(s) &= [a_{k,i,j}(s)]_{i,j}\,, \\
Y_k(s) &= \left[\frac{\lambda_i p_i(k,j)}{s + \lambda_i}\right]_{i,j}, \\
\overline{D}_k(s) &= \left[\int_0^{\infty} e^{-st} P_{i,j}(k,t)(1 - F(t))dt\right]_{i,j}, \\
\overline{A}_k(s) &= \sum_{i=k}^{\infty} A_i(s), \\
B_k(s) &= A_{k+1}(s) - \overline{A}_{k+1}(s)(\overline{A}_0(s))^{-1}, \\
R_0(s) &= \mathbf{0}, \\
R_1(s) &= A_0^{-1}(s), \\
R_k(s) &= R_1(s)[R_{k-1}(s) - \sum_{i=0}^{k-1} A_{i+1}(s)R_{k-i}(s)], \quad k \geq 2, \\
M_b(s) &= R_{b+1}(s)A_0(s) + \sum_{k=0}^{b} R_{b-k}(s)B_k(s) - \sum_{k=b+1}^{\infty} Y_k(s) \\
&\quad - \sum_{k=0}^{b} Y_{b-k}(s)[R_{k+1}(s)A_0(s) + \sum_{i=0}^{k} R_{k-i}(s)B_i(s)].
\end{aligned}$$

and column vectors of size m:

$$\begin{aligned}
\mathbf{1} &= \text{the column vector of 1's}, \\
z(s) &= ((s + \lambda_1)^{-1}, \ldots, (s + \lambda_m)^{-1})^T.
\end{aligned}$$

III. Queue size distribution

In a BMAP queue, all the time-dependent characteristics depend on the initial queue size, $X(0)$, and the initial state of the modulating process, $J(0)$. This dependence will be represented by indices n, i in the queue size distribution:

$$\Phi_{n,i}(t,l) = \mathbf{P}(X(t) = l | X(0) = n, J(0) = i).$$

Naturally, l and n vary from 0 to b while i varies from 1 to m. In the stationary case the dependence on $X(0)$ and $J(0)$ vanishes and we may simply denote the stationary queue size distribution by p_l where

$$p_l = \lim_{t\to\infty} \mathbf{P}(X(t){=}l) = \lim_{t\to\infty} \mathbf{P}(X(t){=}l | X(0){=}n, J(0){=}i)$$

and n, i can be arbitrary.

The main result of this paper is expressed in terms of the Laplace transform:

$$\phi_{n,i}(s,l)=\int_0^\infty e^{-st}\Phi_{n,i}(t,l)dt,$$

and the column vector representing different initial states of the modulating process:

$$\phi_n(s,l)=(\phi_{n,1}(s,l),\ldots,\phi_{n,m}(s,l))^T.$$

Theorem 1: The Laplace transform of the queue size distribution in the $BMAP/G/1/b$ queue has the form:

$$\phi_n(s,l)=\sum_{k=0}^{b-n}R_{b-n-k}(s)g_k(s,l)$$
$$+[R_{b-n+1}(s)A_0(s)+\sum_{k=0}^{b-n}R_{b-n-k}(s)B_k(s)]M_b^{-1}(s)m_b(s,l), \tag{5}$$

where

$$\begin{aligned} g_k(s,l) &= \overline{A}_{k+1}(s)(\overline{A}_0(s))^{-1}r_b(s,l)-r_{b-k}(s,l),\\ r_n(s,l) &= \begin{cases} \mathbf{0}\cdot\mathbf{1}, & \text{if} \quad l<n,\\ \overline{D}_{l-n}(s)\cdot\mathbf{1}, & \text{if} \quad n\le l<b,\\ \frac{1-f(s)}{s}\cdot\mathbf{1}-\sum_{k=0}^{b-n-1}\overline{D}_k(s)\cdot\mathbf{1}, & \text{if} \quad l=b.\end{cases}\\ m_b(s,l) &= \sum_{k=0}^{b}Y_{b-k}(s)\sum_{i=0}^{k}R_{k-i}(s)g_i(s,l)\\ &\quad -\sum_{k=0}^{b}R_{b-k}(s)g_k(s,l)+\delta_{0l}z(s).\end{aligned}$$

Proof. Conditioning on the first departure moment we may write for $0<n\le b$, $1\le i\le m$:

$$\begin{aligned}\Phi_{n,i}(t,l) &= \sum_{j=1}^{m}\sum_{k=0}^{b-n-1}\int_0^t\Phi_{n+k-1,j}(t-u,l)P_{i,j}(k,u)dF(u)\\ &+ \sum_{j=1}^{m}\sum_{k=b-n}^{\infty}\int_0^t\Phi_{b-1,j}(t-u,l)P_{i,j}(k,u)dF(u)\\ &+ \rho_{n,i}(t,l),\end{aligned} \tag{6}$$

where

$$\rho_{n,i}(t,l)=(1-F(t))\cdot\begin{cases}0, & \text{if } l<n,\\ \sum_{j=1}^{m}P_{i,j}(l-n,t), & \text{if } n\le l<b,\\ \sum_{j=1}^{m}\sum_{k=b-n}^{\infty}P_{i,j}(k,t), & \text{if } l=b.\end{cases}$$

Similarly, if $n=0$ then for $1\le i\le m$ we have:

$$\begin{aligned}\Phi_{0,i}(t,l) &= \sum_{j=1}^{m}\sum_{k=0}^{b}\int_0^t\Phi_{k,j}(t-u,l)p_i(k,j)\lambda_ie^{-\lambda_iu}du\\ &+ \sum_{j=1}^{m}\sum_{k=b+1}^{\infty}\int_0^t\Phi_{b,j}(t-u,l)p_i(k,j)\lambda_ie^{-\lambda_iu}du\\ &+ \delta_{0l}e^{-\lambda_it}.\end{aligned} \tag{7}$$

Applying transforms to (6) and (7) yields:

$$\begin{aligned}\phi_{n,i}(s,l) &= \sum_{j=1}^{m}\sum_{k=0}^{b-n-1}a_{k,i,j}(s)\phi_{n+k-1,j}(s,l)\\ &+ \sum_{j=1}^{m}\sum_{k=b-n}^{\infty}a_{k,i,j}(s)\phi_{b-1,j}(s,l)\\ &+ \int_0^\infty e^{-st}\rho_{n,i}(t,l)dt,\end{aligned}$$

and

$$\begin{aligned}\phi_{0,i}(s,l) &= \sum_{j=1}^{m}\sum_{k=0}^{b}p_i(k,j)\phi_{k,j}(s,l)\frac{\lambda_i}{s+\lambda_i}\\ &+ \sum_{j=1}^{m}\sum_{k=b+1}^{\infty}p_i(k,j)\phi_{b,j}(s,l)\frac{\lambda_i}{s+\lambda_i}\\ &+ \delta_{0l}\frac{1}{s+\lambda_i},\end{aligned}$$

respectively. Next, applying matrix notation we get:

$$\begin{aligned}\phi_n(s,l) &= \sum_{k=0}^{b-n-1}A_k(s)\phi_{n+k-1}(s,l)+\sum_{k=b-n}^{\infty}A_k(s)\phi_{b-1}(s,l)\\ &\quad +r_n(s,l), \qquad 0<n\le b,\end{aligned} \tag{8}$$

$$\phi_0(s,l)=\sum_{k=0}^{b}Y_k(s)\phi_k(s,l)+\sum_{k=b+1}^{\infty}Y_k(s)\phi_b(s,l)+\delta_{0l}z(s). \tag{9}$$

Denoting $\varphi_n(s,l)=\phi_{b-n}(s,l)$ we may rewrite (8) and (9) as follows:

$$\sum_{k=-1}^{n}A_{k+1}(s)\varphi_{n-k}(s,l)-\varphi_n(s,l)=\psi_n(s,l), \qquad 0\le n<b, \tag{10}$$

$$\varphi_b(s,l)=\sum_{k=0}^{b}Y_{b-k}(s)\varphi_k(s,l)+\sum_{k=b+1}^{\infty}Y_k(s)\varphi_0(s,l)+\delta_{0l}z(s), \tag{11}$$

where

$$\psi_n(s,l)=A_{n+1}(s)\varphi_0(s,l)-\sum_{k=n+1}^{\infty}A_k(s)\varphi_1(s,l)-r_{b-n}(s,l).$$

Now, the system of equations (10) has the following solution:

$$\varphi_n(s,l)=R_{n+1}(s)c(s,l)+\sum_{k=0}^{n}R_{n-k}(s)\psi_k(s,l), \qquad n\ge 0, \tag{12}$$

where $c(s,l)$ is a function that does not depend on n (see, for comparison, [20], page 343). Therefore we are left with the task of finding $\varphi_0(s,l)$, $\varphi_1(s,l)$, which is necessary for calculating $\psi_k(s,l)$, and the function $c(s,l)$. Substituting $n=0$ in (12) we can easily obtain

$$c(s,l)=A_0(s)\varphi_0(s,l), \tag{13}$$

Substituting $n=0$ in (10) we have

$$\varphi_1(s,l)=(\overline{A_0}(s))^{-1}(\varphi_0(s,l)-r_b(s,l)), \tag{14}$$

which reduces the problem to finding $\varphi_0(s,l)$. Using the boundary condition (11) we get

$$\varphi_0(s,l) = M_b^{-1}(s) m_b(s,l),$$

which finishes the proof. □

Formula (5) may be used in practice in several ways. First, using the well-known limiting behaviour of the Laplace transform we can easily obtain the stationary queue size distribution:

$$p_l = \lim_{t\to\infty} \mathbf{P}(X(t)=l) = \lim_{s\to 0+} s\phi_b(s,l).$$

Instead of b in $\phi_b(s,l)$, any other initial queue size can be chosen. However, using b is recommended as in this case the formula (5) reduces to its simplest form, namely

$$\phi_b(s,l) = M_b^{-1}(s) m_b(s,l).$$

Next, we may obtain the average stationary queue size

$$L = \sum_{k=0}^{b} l p_l$$

and all moments, for instance variance

$$Var = \sum_{k=0}^{b} (l-L)^2 p_l.$$

Furthermore, we can obtain also the time-dependent queue size distribution, average, variance etc. To accomplish that, we have to invert the Laplace transform presented in (5). For instance, in [21] an efficient method based on the Euler summation formula is presented. It has the following form

$$f(t) \approx \sum_{k=0}^{m} \sum_{j=0}^{n+k} \binom{m}{k} 2^{-m} (-1)^j a_j(t), \qquad (15)$$

where

$$a_k(t) = \frac{e^{A/2t}}{2lt} b_k(t), \quad k \geq 0,$$

$$b_0(t) = f^*\left(\frac{A}{2lt}\right) + 2\sum_{j=1}^{l} Re\left[f^*\left(\frac{A}{2lt} + \frac{ij\pi}{lt}\right) e^{ij\pi/t}\right],$$

$$b_k(t) = 2\sum_{j=1}^{l} Re\left[f^*\left(\frac{A}{2lt} + \frac{ij\pi}{lt} + \frac{ik\pi}{t}\right) e^{ij\pi/t}\right], \quad k \geq 1.$$

$f^*(s)$ denotes a transform to be inverted, $f(t)$ is the original function, parameters m, n, A, l are used to control the inversion error. Proposed in [21] typical values are $m = 11$, $n = 38$, $A = 19$ and $l = 1$. An easy verification shows that for this set of control parameters, 51 transform values are required in order to obtain one value of the original function.

IV. Numerical illustration

Before numerical examples are given, it is worth mentioning that matrices $A_k(s)$ and $\overline{D}_k(s)$, which appear in Theorem 1, can be calculated effectively by means of the uniformization technique, described in [2]. For instance, if the service time is constant and equal to d, applying this technique we get:

$$A_n(s) = \sum_{j=0}^{\infty} \gamma_j(s) K_{n,j}, \qquad \overline{D}_n(s) = \sum_{j=0}^{\infty} \delta_j(s) K_{n,j}, \quad (16)$$

where

$$\begin{aligned}
K_{0,0} &= I, \\
K_{n,0} &= \mathbf{0}, \qquad n \geq 1, \\
K_{0,j+1} &= K_{0,j}(I + \theta^{-1} D_0), \\
K_{n,j+1} &= \theta^{-1} \sum_{i=0}^{n-1} K_{i,j} D_{n-i} + K_{n,j}(I + \theta^{-1} D_0), \\
\theta &= \max_i \{(-D_0)_{ii}\}, \\
\gamma_j(s) &= \frac{e^{-(\theta+s)d} (\theta d)^j}{j!}, \\
\delta_j(s) &= \theta^j \frac{\Gamma(j+1,0) - \Gamma(j+1, d(s+\theta))}{j!(s+\theta)^{j+1}},
\end{aligned}$$

and $\Gamma(j,x)$ denotes the incomplete gamma function.

The remaining matrices and vectors in Theorem 1 are either trivial (like $Y_k(s)$, $z(s)$) or simple functions of $A_k(s)$ and $\overline{D}_k(s)$.

A. Example 1

In this example we use three different BMAPs to demonstrate an impact of the autocorrelation in the arrival process on the queue size distribution. Namely, we assume that

$$D_0 = \begin{bmatrix} -10 & 0.1 \\ 0.01 & -0.1 \end{bmatrix},$$

is common and set

1) $BMAP_0$:

$$D_1 = \begin{bmatrix} 0.909367 & 0.080633 \\ 0.008267 & 0.000733 \end{bmatrix},$$

$$D_2 = \begin{bmatrix} 1.818734 & 0.161266 \\ 0.016534 & 0.001466 \end{bmatrix},$$

$$D_5 = \begin{bmatrix} 6.365569 & 0.564431 \\ 0.057869 & 0.005131 \end{bmatrix},$$

2) $BMAP_{0.2}$:

$$D_1 = \begin{bmatrix} 0.947856 & 0.042144 \\ 0.004332 & 0.004668 \end{bmatrix},$$

$$D_2 = \begin{bmatrix} 1.895711 & 0.084289 \\ 0.008663 & 0.009337 \end{bmatrix},$$

$$D_5 = \begin{bmatrix} 6.634989 & 0.295011 \\ 0.030321 & 0.032679 \end{bmatrix},$$

3) $BMAP_{0.4}$:

$$D_1 = \begin{bmatrix} 0.986344 & 0.003656 \\ 0.000396 & 0.008604 \end{bmatrix},$$

$$D_2 = \begin{bmatrix} 1.972688 & 0.007312 \\ 0.000793 & 0.017207 \end{bmatrix},$$

$$D_5 = \begin{bmatrix} 6.904409 & 0.025591 \\ 0.002774 & 0.060226 \end{bmatrix},$$

These parameterizations were chosen in such a way that the resulting BMAPs have common average batch size $= 4$, common batch arrival rate $\Lambda_g = 1$, common total arrival rate $\Lambda = 4$ and common variance of the interarrival time $Var = 17.21$. However, the correlations between two consecutive interarrival times (see (4)) are equal to 0, 0.2 and 0.4 for $BMAP_0$, $BMAP_{0.2}$ and $BMAP_{0.4}$, respectively.

It is assumed that the service time is constant and equal to 0.2, which makes the load offered to the queue to be of 80%. The buffer size is 50.

We can now obtain sample queue size distributions. We start with transient distribution, which depends on the initial state of the system. We assume that initially the buffer is full ($X(0) = b = 50$) and that the initial state of the modulating process is 1.

Figures 1-4 show transient queue size distributions in times $t = 2.5$, $t = 5$, $t = 10$ and $t = 20$, respectively. In each figure three curves, representing $BMAP_0$, $BMAP_{0.2}$ and $BMAP_{0.4}$, are depicted.

Firstly, we can see that these distributions may assume rather complicated shapes. This effect is typical for the batch arrival queue and the irregularities in the shape are connected with batch sizes and their combinations. In this example we have three possible batch sizes (1, 2, 5), therefore we can observe concentrations of probability mass connected with arrivals of 1, 2, 5 cells, but also with 1+2, 1+5, 2+5 etc.

Secondly, the autocorrelation in the arrival process severely influences the queue size distribution. The higher the autocorrelation, the more of the probability mass concentrated around b and the higher the full-buffer probability, which is connected with losses.

What is more, the autocorrelation in the arrival process influences the convergence to the steady-state distribution. This effect can be observed in Figs. 5-7. For $BMAP_0$, distribution closely resembling the shape of the steady-state curve is obtained for t around 15. For $BMAP_{0.2}$ it takes around 20 to obtain steady state while for $BMAP_{0.4}$ the steady-state curve can be obtained for t over 30.

B. Example 2

In this example we demonstrate the queue size distribution using a BMAP parameterization based on measurements of aggregated IP traffic. For this purpose, a trace file recorded at the Front Range GigaPOP (FRG) aggregation point, which is run by PMA (Passive Measurement and Analysis Project, see http://pma.nlanr.net) has been utilized[1]. The average rate of the traffic is 72 MBytes/s, with mean packet size of 869Bytes.

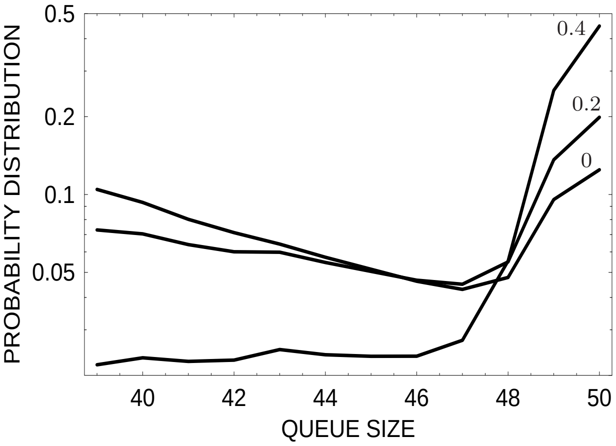

Fig. 1. Transient queue size distributions at time $t = 2.5$ for $BMAP_0$, $BMAP_{0.2}$ and $BMAP_{0.4}$ arrivals (Example 1).

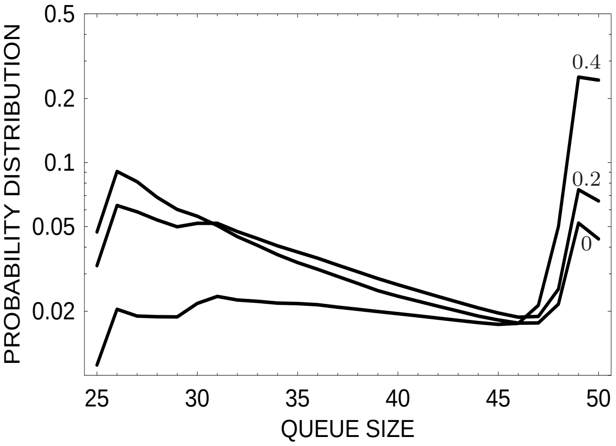

Fig. 2. Transient queue size distributions at time $t = 5$ for $BMAP_0$, $BMAP_{0.2}$ and $BMAP_{0.4}$ arrivals (Example 1).

As IP traces are often dominated by several most frequent packet sizes [10], it is not necessary to take all possible packet sizes into account. In the sample used herein, seven packet sizes (40, 52, 552, 1300, 1420, 1488, 1500) account for 97 percent of the traffic and only these sizes were used. Using the expectation-maximization (EM) method [10] the following BMAP parameters were estimated:

$$D_0 = \begin{bmatrix} -90020.6 & 5300.2 & 11454.4 \\ 9132.5 & -126814.5 & 14807.8 \\ 2923.1 & 198.6 & -94942.6 \end{bmatrix},$$

$$D_{40} = \begin{bmatrix} 2898.0 & 3415.6 & 1365.3 \\ 3649.6 & 1510.9 & 1044.6 \\ 1943.3 & 1954.9 & 7696.6 \end{bmatrix},$$

$$D_{52} = \begin{bmatrix} 10980.1 & 8180.1 & 3284.0 \\ 5875.6 & 19512.8 & 19443.1 \\ 4064.1 & 9956.0 & 2744.3 \end{bmatrix},$$

$$D_{552} = \begin{bmatrix} 1259.3 & 1143.0 & 960.4 \\ 1343.6 & 1541.2 & 3903.5 \\ 2470.5 & 1607.7 & 665.2 \end{bmatrix},$$

[1] Precisely, one million packet headers from the trace file FRG-1137458803-1.tsh, recorded on Jan 17th, 2006, were used.

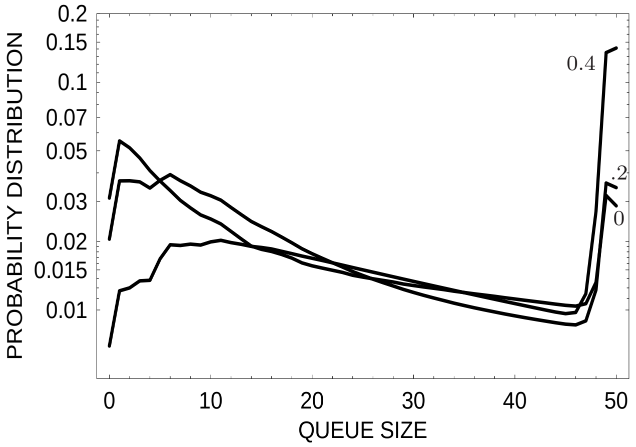

Fig. 3. Transient queue size distributions at time $t = 10$ for $BMAP_0$, $BMAP_{0.2}$ and $BMAP_{0.4}$ arrivals (Example 1).

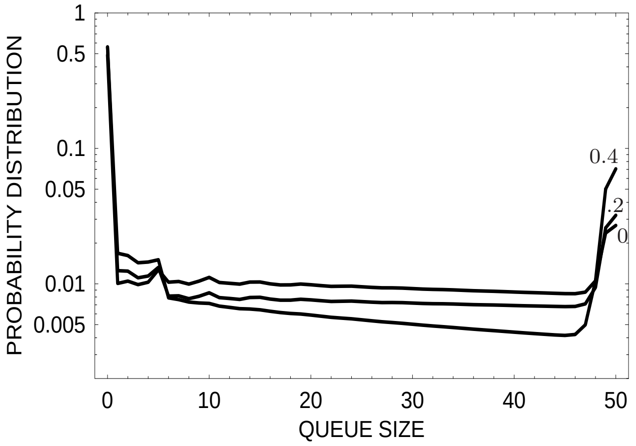

Fig. 4. Transient queue size distributions at time $t = 20$ for $BMAP_0$, $BMAP_{0.2}$ and $BMAP_{0.4}$ arrivals (Example 1).

$$
\begin{aligned}
D_{1300} &= \begin{bmatrix} 115.3 & 174.5 & 188.1 \\ 333.8 & 28.0 & 405.9 \\ 337.3 & 124.4 & 552.4 \end{bmatrix}, \\
D_{1420} &= \begin{bmatrix} 878.9 & 65.6 & 506.4 \\ 916.4 & 31.1 & 999.1 \\ 205.0 & 138.1 & 1603.7 \end{bmatrix}, \\
D_{1488} &= \begin{bmatrix} 95.1 & 93.2 & 61.3 \\ 199.5 & 175.3 & 129.7 \\ 192.2 & 34.6 & 107.5 \end{bmatrix}, \\
D_{1500} &= \begin{bmatrix} 13997.0 & 14089.9 & 9514.9 \\ 31145.3 & 9632.8 & 1052.4 \\ 17681.9 & 14814.8 & 22926.4 \end{bmatrix}.
\end{aligned}
$$

Basic characteristics of the original traffic sample and its BMAP model are shown in Table I.

In practice, packets are usually segmented into fixed size units prior to storage, which is connected with memory architecture for fast packet buffering in routers and switches [23]. Therefore, for computation of the buffer occupancy distribution we will rather use the following BMAP parameters:

$$D'_1 = D_{40} + D_{52}, \quad D'_9 = D_{552}, D'_{21} = D_{1300},$$

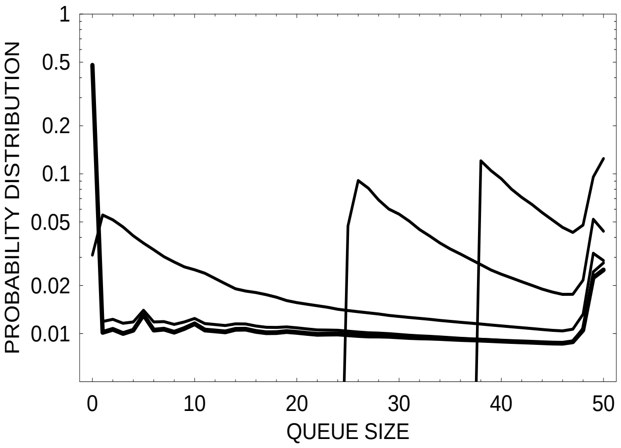

Fig. 5. Transient queue size distributions for $BMAP_0$ arrivals (Example 1). Each curve represents a different moment in time, namely $t = 2.5$, $t = 5$, $t = 10$, $t = 15$, counting from the top. The lowest, thick curve represents steady state.

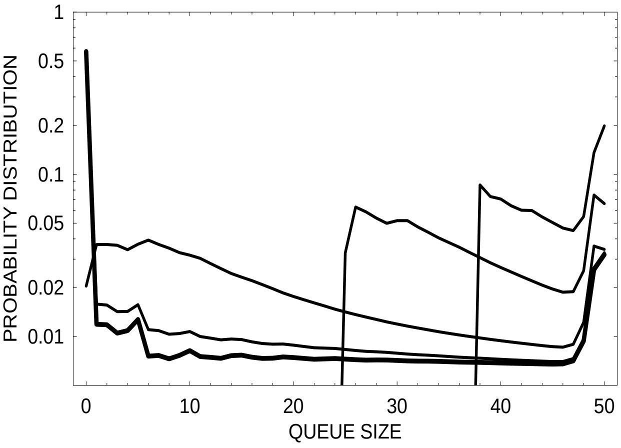

Fig. 6. Transient queue size distributions for $BMAP_{0.2}$ arrivals (Example 1). Each curve represents a different moment in time, namely $t = 2.5$, $t = 5$, $t = 10$, $t = 15$, counting from the top. The lowest, thick curve represents steady state.

$$D'_{23} = D_{1420}, \quad D'_{24} = D_{1488} + D_{1500}.$$

where the new indices denote numbers of occupied units, assuming typical unit size of 64Bytes [23].

We assume that the queue is served at the constant rate of 80MB/s (1310720 units/s) and the buffer size is 100KBytes. The initial state of the modulating process, $J(0)$, is distributed according to

$$\pi = (0.39517, 0.24563, 0.35920). \tag{17}$$

Now we are in a position to present the numerical results. In Figs. 8, 9 the stationary queue size distribution for the system considered is presented. In particular, in Fig 8 the whole distribution is depicted while in 9 a close-up of range 0-5KB is shown. The average queue size is:

$$L = 9.584 \text{ KBytes},$$

while its standard deviation:

$$\sqrt{Var} = 9.998 \text{ KBytes},$$

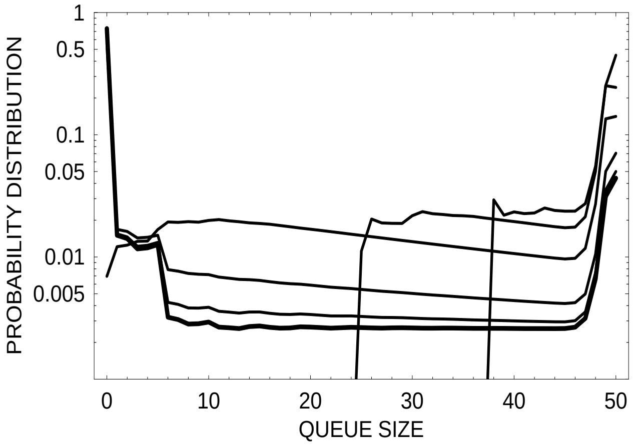

Fig. 7. Transient queue size distributions for $BMAP_{0.4}$ arrivals (Example 1). Each curve represents a different moment in time, namely $t = 2.5$, $t = 5$, $t = 10$, $t = 20$, $t = 30$, counting from the top. The lowest, thick curve represents steady state.

	traffic sample	BMAP
mean packet interarr. time [μs]	11.467	11.467
standard dev. of the interarr. time [μs]	11.599	11.594
mean packet size [Bytes]	869.18	869.38
total arrival rate [MBytes/s]	72.286	72.301

TABLE I
PARAMETERS OF THE ORIGINAL TRAFFIC SAMPLE AND ITS BMAP MODEL.

The probability that the buffer is empty equals to

$$p_0 = 6.769 \times 10^{-2},$$

while the probability that the buffer is full is

$$p_b = 1.367 \times 10^{-7}.$$

As in Example 1, it is striking that the distribution has jumps in the lower range and its shape is rather complicated. Again, this effect is caused by the batch structure of the arrival process and in a way reflects the batch size distribution. For instance, the high peak observed in Fig. 9 around 1.5KB is connected with high probability of 1500B packet and high values of elements of D_{1500}.

On the other hand, for larger queue sizes the shape becomes simple and the function is approximately linear (on a log-scaled plot).

The statistical structure of the BMAP, which reflects the structure of the original traffic, has a deep impact on the queue size distribution. To demonstrate this, we may replace the BMAP by the Poisson process with exactly the same arrival rate. In this case we obtain [2]

$$\begin{aligned} L &= 470.6 \text{ Bytes}, \\ \sqrt{Var} &= 464.0 \text{ Bytes}, \\ p_b &= 3.792 \times 10^{-98}. \end{aligned}$$

This set of numbers, which is in dramatic contrast to the previous one, illustrates how misleading it may be to neglect taking the precise statistical structure of the traffic into account.

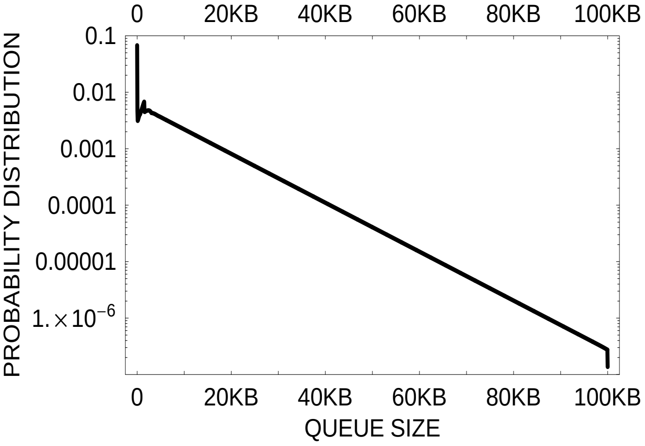

Fig. 8. Stationary queue size distribution in Example 2.

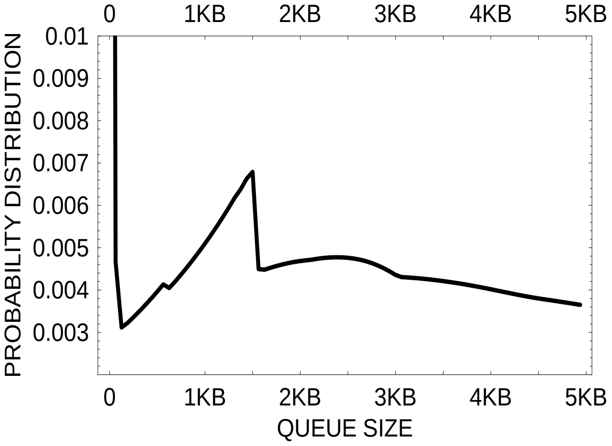

Fig. 9. Stationary queue size distribution in Example 2 - a close-up of the range 0-5KB.

V. Conclusions

In this paper, an analysis of the queue size distribution in a BMAP queue with finite buffer was conducted. It is reasonable to believe that the presented results are of practical importance due to the following reasons. Firstly, a very flexible arrival process, which among other things can model IP traffic, was considered. Secondly, the finite buffer was assumed. In a real network all elements (switches, routers etc.) have finite buffers, which causes losses and influences the network performance. Thirdly, the results were presented in a closed, easy to use form, and they permit one to obtain both, transient and stationary queue size distributions. Finally, computational remarks were given. In particular, formulas for the numerical calculation of the coefficient matrices $A_k(s)$ and $\overline{D}_k(s)$, which occur in the main theorem, were shown.

[2] The formulae for the classic $M/G/1/b$ queue, which were used for obtaining this set of results, can be found by the reader in [22], page 202.

ACKNOWLEDGEMENT

This material is based in part upon work supported by the Polish Ministry of Scientific Research and Information Technology under Grant No. 3 T11C 014 26.

REFERENCES

[1] Neuts, M. F. A Versatile Markovian Point Process, Journal of Applied Probability 14, 764-779, (1979).

[2] Lucantoni, D. M. New results on the single server queue with a batch Markovian arrival process. Commun. Stat., Stochastic Models 7, No.1, 1-46 (1991).

[3] Leland, W., Taqqu, M., Willinger, W. and Wilson, D. On the self-similar nature of ethernet traffic (extended version), IEEE/ACMTransactions on Networking 2(1): 115, (1994).

[4] Crovella, M. and Bestavros, A. Self-similarity in World Wide Web traffic: Evidence and possible causes, IEEE/ACM Transactions on Networking 5(6): 835-846, (1997).

[5] Shah-Heydari, S. and Le-Ngoc, T. MMPP models for multimedia traffic. Telecommunication Systems 15, No.3-4, 273-293 (2000).

[6] Wong, T. C., Mark, J. W. Chua, K. C. Delay performance of voice and MMPP video traffic in a cellular wireless ATM network. IEE Proceedings Communications, Vol. 148, 302-309 (2001).

[7] Wu, G. L. and Mark, J. W. Computational Methods for Performance Evaluation of a Statistical Multiplexer Supporting Bursty Traffic. IEEE/ACM Transactions on Networking, Vol. 4, No. 3, pp. 386-397, (1996).

[8] Kim, Y. H. and Un, C. K. Performance analysis of statistical multiplexing for heterogeneous bursty traffic in ATM network. IEEE Trans. Commun. 42(2-4) pp. 745-753, (1994).

[9] Skelly, P., Schwartz, M. and Dixit, S. A histogram-based model for video traffic behavior in an ATM multiplexer. IEEE/ACM Trans. Netw. 1(4): 446-459 (1993).

[10] Klemm, A., Lindemann, C. and Lohmann, M. Modeling IP traffic using the batch Markovian arrival process. Performance Evaluation, Vol. 54, Issue 2, (2003).

[11] Salvador, P., Pacheco, A. and Valadas, R. Modeling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs. Computer Networks 44, pp. 335-352, (2004).

[12] Landman, J. and Kritzinger, P. Delay analysis of downlink IP traffic on UMTS mobile networks. Performance Evaluation 62, pp. 68-82 (2005).

[13] Klemm, A., Lindemann, C. and Lohmann, M. Traffic Modeling and Characterization for UMTS Networks, Proc. GLOBECOM 2001, pp. 1741–1746, San Antonio TX, November (2001).

[14] Ramaswami, V. The $N/G/1$ queue and its detailed analysis. Adv. Appl. Prob. 12, 222-261 (1980).

[15] Lucantoni, D. M., Choudhury, G. L. and Whitt, W. The transient $BMAP/G/1$ queue. Commun. Stat., Stochastic Models 10, No.1, 145-182 (1994).

[16] Lucantoni, D. Further transient analysis of the $BMAP/G/1$ queue. Commun. Stat., Stochastic Models 14, No.1-2, 461-478 (1998).

[17] Blondia, C. The finite Capacity $N/G/1$ Queue. Communications in Statistics: Stochastic Models, 5(2):273–294, (1989).

[18] Baiocchi, A. and Blefari-Melazzi, N. Steady-state analysis of the MMPP/G/1/K queue. IEEE Trans. Commun. 41, No.4, 531-534 (1992).

[19] http://www.omnetpp.org/

[20] Chydzinski, A. The oscillating queue with finite buffer. Performance Evaluation, 57/3 pp. 341-355, (2004).

[21] J. Abate, G. L. Choudhury and W. Whitt. An introduction to numerical transform inversion and its application to probability models. Chapter in Computational Probability, pp. 257-323, W. Grassman (ed.), Kluwer, Boston, 1999.

[22] Takagi, H. *Queueing analysis. Vol. 2. Finite Systems*. North-Holland, Amsterdam. (1993).

[23] Iyer, S., Kompella, R. R. and McKeown, N. Analysis of a memory architecture for fast packet buffers. In Proc. of IEEE High Performance Switching and Routing, Dallas, Texas, May (2001).

[24] Chydzinski, A. Queue Size in a BMAP Queue with Finite Buffer. Lecture Notes in Computer Science 4003, pp. 200-210, Proc. of NEW2AN'06, St. Petersburg, May 2006.

Andrzej Chydzinski received his MS (in mathematics) and PhD (in computer science) degrees from Silesian University of Technology, Gliwice, Poland, in 1997 and 2002, respectively. He is currently an Assistant Professor in the Institute of Computer Sciences of this university. His academic and professional interests include mathematical modeling and simulation of queueing systems and computer networks. Dr. Chydzinski has an established record of publications in these fields, with more than forty technical journal and conference published papers.