Operations Research Letters 45 (2017) 467–470

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Computation of the moments of queue length in the *BMAP*/*SM*/1 queue

V. Klimenok [a,b], A.N. Dudin [a,b,*], K. Samouylov [b]

[a] Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus
[b] Department of Applied Probability and Informatics, RUDN University, 6, Miklukho-Maklaya st., 117198, Moscow, Russia

ARTICLE INFO

Article history:
Received 31 March 2017
Received in revised form 9 June 2017
Accepted 15 July 2017
Available online 27 July 2017

Keywords:
Batch Markovian arrival process
Semi-Markovian service
Moments of queue length

ABSTRACT

The *BMAP*/*SM*/1 queue is the most general single-server queueing model which can be analysed analytically. Problem of computation of stationary distributions of queue length is solved in the literature. However, the problem of computation of the moments of these distributions is not enough addressed. This problem is more complicated than its particular case when the service times are independent identically distributed random variables due to reducibility of some involved matrices. In this communication, we solve this problem.

1. Introduction

The *BMAP*/*G*/1 queue as a single server system with infinite buffer, Batch Markovian Arrival Process (*BMAP*) and independent identically arbitrarily distributed service time is very important for applications queueing model because the *BMAP* (see [2,11]) is an adequate mathematical model of bursty, correlated flows of information in modern telecommunication networks. Therefore its analysis was very important theoretical task. Initially, this task was solved long time ago by V. Ramaswami in [14] where essentially the same arrival process as the *BMAP* was called as *N* process. The *BMAP*/*G*/1 queue was then analysed in [11]. However, until now analysis of such a system attracts attention of researchers, see, e.g., very recent paper [17].

As it was mentioned above, the *BMAP*/*G*/1 type model assumes independence and identical distribution of service times of successive customers. In many real-world systems, service times of successive customers may be dependent and have different distributions. To take into account such a dependence, so called *Semi-Markovian* (*SM*) service process was considered, see, e.g., [3,18]. Importance of consideration of queues with *SM* service process for practical needs is stressed, e.g., in recent work [1] and references therein. In the paper [16], a very general model of the *SM*/*SM*/1 type with possible dependence of inter-arrival and service times was considered under assumption that the marginal distribution of service times is of phase type. The *BMAP*/*SM*/1 type queue without such an assumption and with batch arrivals was analysed in [7,12]. The problem of computation of steady state distribution of queue length and waiting time distribution in the system was successfully solved. In this communication, we supplement results of [12] by the effective recursive procedures for computation of the moments of the queue length distributions. Moments have an important role for performance evaluation of various queueing systems. Sometimes, information about the mean value and variance of queue length is enough for managerial decisions. If this information is insufficient and the shape of queue length distribution is of a primary interest (e.g., to evaluate the probability that the queue length will exceed a certain important level) while this shape hardly can be found exactly, the shape can be estimated numerically based on the knowledge of the value of several moments of the distribution, see, e.g. [19]. Effective recursive procedures for computing the moments of the queue length at service completion epochs and arbitrary time for *BMAP*/*G*/1 queue were given in [6]. Direct extension of results from [6] appears not possible for the *BMAP*/*SM*/1 type queue because that results essentially exploit irreducibility of some matrix generating functions at the point $z = 1$ while they are reducible when the service is of *SM* type. In this communication we elaborate the recursive procedures for computation of the moments of the queue length in this *BMAP*/*SM*/1 type system.

2. Preliminary results

We consider a single server system with an infinite buffer. The arrival process is the *BMAP*. It is defined by the underlying process $\nu_t,\ t \geq 0$, which is an irreducible continuous time Markov chain

* Corresponding author at: Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus.

E-mail addresses: klimenok@bsu.by (V. Klimenok), dudin@bsu.by (A.N. Dudin), ksam@sci.pfu.edu.ru (K. Samouylov).

http://dx.doi.org/10.1016/j.orl.2017.07.003
0167-6377/

with a finite state space $\{0, \ldots, W\}$ and with the matrix generating function $D(z) = \sum_{k=0}^{\infty} D_k z^k,\ |z| \le 1$, of square matrices $D_k,\ k \ge 0$, of size $(W+1)$ consisting of the intensities of transitions of the Markov chain ν_t accompanied by the generation of k-size batch of customers, $k \ge 0$. The matrix $D(1)$ is an infinitesimal generator of the process ν_t. The stationary distribution vector $\boldsymbol{\theta}$ of this process is the unique solution of the system $\boldsymbol{\theta} D(1) = \boldsymbol{0},\ \boldsymbol{\theta}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector consisting of 1's, and $\boldsymbol{0}$ is a row vector of 0's. The average intensity λ (fundamental rate) of the *BMAP* is given by $\lambda = \boldsymbol{\theta} D'(z)|_{z=1}\mathbf{e}$. We assume that $\lambda < \infty$. For more detailed and exact definition of the *BMAP* see [2,11].

The successive service times of customers are defined as the sojourn times of the semi-Markovian process $m_t,\ t \ge 0$, (see [3]) in its states. This process has the finite state space $\{1, \ldots, M\}$ and the semi-Markovian kernel $\boldsymbol{B}(t) = (B_{m,m'}(t))_{m,m'=\overline{1,M}}$. Here and in the sequel notation like $m = \overline{1, M}$ means that the variable m admits the values from the set $\{1, \ldots, M\}$. Matrix $\boldsymbol{B}(\infty)$ is assumed to be irreducible. The average service time is calculated as $b_1 = \boldsymbol{b} \int_0^{\infty} t d\boldsymbol{B}(t)\mathbf{e}$ where the row-vector $\boldsymbol{b}$ is a vector of the stationary distribution of the embedded Markov chain for the semi-Markovian process m_t. This vector is computed as the unique solution to the system $\boldsymbol{b}\boldsymbol{B}(\infty) = \boldsymbol{b},\ \boldsymbol{b}\mathbf{e} = 1$.

Let $i_t,\ \ i_t \ge 0$, be the queue length at the moment $t \ge 0$. The process i_t is non-Markovian. To study this process, first we consider the three-dimensional process $\{i_t, m_t, \nu_t\},\ \ t \ge 0$, and then consider this process only at the instances of service completion. Let $i_n,\ \ i_n \ge 0$, be the number of the customers in the system, $\nu_n,\ \ \nu_n = \overline{0, W}$, be the state of the *BMAP* underlying process and $m_n,\ \ m_n = \overline{1, M}$, be the state of the *SM* underlying process immediately after the n th service completion instant in the system, $n \ge 1$. It is easy to see that the three-dimensional process $\xi_n = \{i_n, m_n, \nu_n\},\ n \ge 1$, is a discrete-time Markov chain (embedded Markov chain for the considered queueing model). This Markov chain belongs to the class of $M/G/1$ type Markov chains, see [13].

Remark 1. In paper [12], the finite components, m_n and ν_n, of the Markov chain $\xi_n,\ \ n \ge 1$, are listed in another order. In principle, the order of these components is not very important. However, from the algorithmic point of view, the choice made above is better due to two reasons: (a) possibility of the direct use of the known procedures for computation of the matrices $P(n,t)$, entries of which define probability to have n arrivals in the *BMAP* during time t (otherwise, a tiresome procedure consisting of many sequential coordinated permutations of rows and columns is required, see [5]); (b) a reducible matrix $\tilde{\boldsymbol{D}}$ appearing below (in Theorem 2) already has the required for the algorithmic purposes canonical normal form.

Introduce the vector generating functions

$$Y(z) = \boldsymbol{\beta}(-D(z)) = \int_0^{\infty} d\boldsymbol{B}(t) \otimes e^{D(z)t},$$

$$V(z) = \frac{1}{z}(-\tilde{D}_0)^{-1}(\tilde{D}(z) - \tilde{D}_0)\boldsymbol{\beta}(-D(z))$$

where $\tilde{D}_0 = I_M \otimes D_0,\ \ \tilde{D}(z) = I_M \otimes D(z)$, $\otimes$ is symbol of Kronecker product of matrices, see [9].

Let also $\boldsymbol{A}(z) = zI - Y(z)$.

As follows, e.g., from [8], the necessary and sufficient condition for ergodicity of the $M/G/1$ type Markov chain $\xi_n,\ \ n \ge 1$, is the fulfilment of the inequality

$$\gamma = [det\boldsymbol{A}(z)]'|_{z=1} > 0. \tag{1}$$

In our case, this inequality reduces to the inequality $\rho < 1$ where the parameter $\rho = \lambda b_1$ is called as the load of the system. In what follows we assume that this condition holds. Then the stationary probabilities

$$p(i, m, \nu) = \lim_{t \to \infty} P\{i_t = i,\ m_t = m,\ \nu_t = \nu\},$$

$$\pi(i, m, \nu) = \lim_{n \to \infty} P\{i_n = i,\ m_n = m,\ \nu_n = \nu\}, i \ge 0,$$

$$m = \overline{1, M},\ \ \nu = \overline{0, W},$$

exist.

Let $\boldsymbol{p}_i,\ \ \boldsymbol{\pi}_i,\ i \ge 0$, be the row vectors formed by the probabilities $p(i, m, \nu)$ and $\pi(i, m, \nu)$ enumerated in the lexicographic order of the components (m, ν).

Introduce the vector generating functions

$$\boldsymbol{P}(z) = \sum_{i=0}^{\infty} \boldsymbol{p}_i z^i,\ \ \boldsymbol{\Pi}(z) = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i z^i,\ \ |z| \le 1.$$

The following results are known, see, e.g., [12].

Proposition 1. *The vector generating function $\boldsymbol{\Pi}(z)$ is the unique analytical solution to the vector functional equation*

$$\boldsymbol{\Pi}(z)(zI - \boldsymbol{\beta}(-D(z)))$$
$$= \boldsymbol{\Pi}(0)(-\tilde{D}_0)^{-1}\tilde{D}(z)\boldsymbol{\beta}(-D(z)) \tag{2}$$

satisfying normalization condition $\boldsymbol{\Pi}(1)\mathbf{e} = 1$.

Proposition 2. *The vector generating functions $\boldsymbol{P}(z)$ and $\boldsymbol{\Pi}(z)$ are related as follows:*

$$\boldsymbol{P}(z) = \lambda \boldsymbol{\Pi}(z)(z\boldsymbol{\beta}^{-1}(-D(z))\nabla^*(z) - I)(\tilde{D}(z))^{-1} \tag{3}$$

where

$$\nabla^*(z) = \int_0^{\infty} d\nabla_{\boldsymbol{B}}(t) \otimes e^{D(z)t},$$

$$\nabla_{\boldsymbol{B}}(t) = \mathrm{diag}\left\{\sum_{m'=1}^{M} B_{m,m'}(t),\ \ m = \overline{1, M}\right\},$$

$\mathrm{diag}\{\ldots\}$ *denotes the diagonal matrix with the diagonal entries listed in the brackets.*

In the literature, there exist several well-known approaches for solving the equations of type (2), in particular, approach by M. Neuts, see, e.g. [13,14] and the transform approach. In the transform approach, see, e.g., [4,8], the unknown vector $\boldsymbol{\Pi}(0)$ is computed by exploiting the analyticity property of the vector generating function $\boldsymbol{\Pi}(z)$ in the unit disk $|z| < 1$ of the complex plane.

Having the vector $\boldsymbol{\Pi}(0) = \boldsymbol{\pi}_0$ been computed, the problem of computation of the vector generating function $\boldsymbol{\Pi}(z),\ \ |z| < 1$, can be considered solved. If, for some purpose, the values of some of the vectors $\boldsymbol{\pi}_i,\ i \ge 1$, are interesting, then these vectors can be recursively computed from equilibrium equations or via the numerically stable procedures from [10,15].

Quite often, it is more important to compute not the probability vectors $\boldsymbol{\pi}_i,\ i \ge 0$, but the moments of the distribution (e.g., the average value and variance of queue length, etc.). Formulas for computation of the initial moments are trivial under known vectors $\boldsymbol{\pi}_i,\ 0 \ge 1$. E.g., the vector moment $\hat{\boldsymbol{M}}_k$ of order k of the distribution of the embedded Markov chain ξ_n is computed as $\hat{\boldsymbol{M}}_k = \sum_{i=1}^{\infty} i^k \boldsymbol{\pi}_i,\ \ k \ge 0$. Computations based on this formula can be easily implemented if the load of the system ρ is not high. If it is high, say $\rho > 0.8$, essential problems can arise. Therefore, it is necessary to have an alternative way for computation of the moments. It is well known that the moments $\hat{\boldsymbol{M}}_k$ can be easy computed if the