Int. J. Appl. Math. Comput. Sci., 2017, Vol. 27, No. 1, 119–131
DOI: 10.1515/amcs-2017-0009

amcs

ANALYSIS OF AN MAP/PH/1 QUEUE WITH FLEXIBLE GROUP SERVICE

ARIANNA BRUGNO [a], CIRO D'APICE [a], ALEXANDER DUDIN [b], ROSANNA MANZO [a,*]

[a] Department of Information Engineering, Electrical Engineering and Applied Mathematics
University of Salerno, Via Giovanni Paolo II, 132, 84084, Fisciano (SA), Italy
e-mail: {abrugno,cdapice,rmanzo}@unisa.it

[b] Department of Applied Mathematics and Computer Science
Belarusian State University, 4, Nezavisimosti Av., Minsk, 220030, Belarus
e-mail: dudin@bsu.by

A novel customer batch service discipline for a single server queue is introduced and analyzed. Service to customers is offered in batches of a certain size. If the number of customers in the system at the service completion moment is less than this size, the server does not start the next service until the number of customers in the system reaches this size or a random limitation of the idle time of the server expires, whichever occurs first. Customers arrive according to a Markovian arrival process. An individual customer's service time has a phase-type distribution. The service time of a batch is defined as the maximum of the individual service times of the customers which form the batch. The dynamics of such a system are described by a multi-dimensional Markov chain. An ergodicity condition for this Markov chain is derived, a stationary probability distribution of the states is computed, and formulas for the main performance measures of the system are provided. The Laplace–Stieltjes transform of the waiting time is obtained. Results are numerically illustrated.

Keywords: queueing system, batch service, multi-rate service, stationary distribution, optimization.

1. Introduction

An overwhelming majority of the queueing literature is devoted to queues where service to customers is provided one by one. However, queues with batch (bulk, group) service also receive their portion of attention. In such queues, service is provided not to an individual customer, but to groups of customers. Usually, the minimal and the maximal size of a group are predefined. Among papers dealing with a batch service discipline, the ones by Bailey (1954), Deb and Serfozo (1973), Downton (1955), Neuts (1967), as well as the survey by Sasikala and Indhira (2016), are representative.

Recent results for queues with group service have been obtained by Banerjee *et al*. (2015). There, many examples of real world applications of queues with group service are given and a survey of related research is presented. According to Banerjee *et al*. (2015), group service queueing systems can be divided into the following categories:

1. systems in which the buffer size is (a) finite or (b) infinite,
2. systems in which arrivals occur according to (a) a renewal or (b) a correlated process,
3. systems in which arrivals occur (a) singly or (b) in batches,
4. systems in which services are (a) independent of the batch size or (b) dependent on the size of the batch being served,
5. systems in which service times are (a) exponential or (b) non-exponential.

It is stressed by Banerjee *et al*. (2015) that very few papers deal with Case 2(b) in combination with 4(b) and 5(b). Our paper deals precisely with this combination. The model considered has two distinguishing features:

- It is usually assumed in the analysis of group service queueing systems that some integer threshold, say N, is fixed and service does not start if the number of

*Corresponding author

customers in the queue is less than N. In our paper, we assume that the idle time of the server is limited, and if this time expires then service starts, even if the number of customers in the queue is less than N. Our assumption better suits certain real world situations. For example, in modelling operation of a passenger delivering system at an airport, the shuttle has to start travel even if it is not completely loaded because (i) the passenger can be late to his/her flight due to a long waiting time and (ii) there is some schedule and the next shuttle should arrive for loading.

- Although a majority of the obtained analytical results are valid for an arbitrary dependence of the batch service time on the number of customers in the batch being served, in our numerical examples we assume that the service time of a batch is defined as the maximum of individual service times of the customers which form a batch. This assumption comes from the model of the so-called multi-rate information transmission that is assumed, e.g., in IEEE802.11 WLAN.

The rest of the paper is organized as follows. In Section 2, the mathematical model is described. The process of system states is introduced in Section 3 as a continuous-time five-dimensional quasi-birth-and-death process and its generator is described. In Section 4, the ergodicity condition is derived. In Section 5, an algorithm for computing stationary distributions of the system states and expressions for key performance indices of the system are presented. The Laplace–Stieltjes transform (LST) of the waiting time distribution is obtained in Section 6. Results of numerical experiments are given in Section 7. They show the advantage of the customer service discipline considered over the classical discipline without the possibility of starting service before the number of customers in the system reaches a predefined threshold value. Section 8 concludes the paper.

2. Mathematical model

We consider a single server queueing system, in which the input flow is described by a Markovian arrival process (MAP). Customer arrival in the MAP is directed by an underlying irreducible continuous time Markov chain ν_t, $t \geq 0$, with a finite state space $\{0, \dots, W\}$. The sojourn time of the Markov chain ν_t, $t \geq 0$, in the state ν has the exponential distribution with parameter $\lambda_\nu, \nu \in \{0, \dots, W\}$. From now on we will use the notation of the type $\nu = \overline{0, W}$ to indicate that ν takes values from the set $\{0, \dots, W\}$. After this sojourn time expires, with probability $p_k(\nu, \nu')$, the process ν_t jumps to the state ν', and k customers, $k = 0, 1$, arrive into the system. The intensities of jumps of the underlying Markov chain from one state into another with generation of k customers are combined into the matrices D_k, $k = 0, 1$, of size $(W+1) \times (W+1)$. The matrix $D(1) = D_0 + D_1$ is the infinitesimal generator of the process ν_t, $t \geq 0$. The invariant probability vector (stationary distribution vector) $\boldsymbol{\theta}$ of this process is computed as the unique solution to the equations

$$\boldsymbol{\theta} D(1) = \mathbf{0}, \quad \boldsymbol{\theta}\mathbf{e} = 1.$$

Here and in the sequel $\mathbf{0}$ is the zero row vector and $\mathbf{e}$ is the column vector of the appropriate size consisting of ones. If the dimensionality of the vector is not clear from the context, it is indicated as a lower index, e.g., $\mathbf{e}_{\overline{W}}$ denotes the unit column vector of dimensionality $\overline{W} = W + 1$. The average intensity λ (fundamental rate) of the MAP is defined as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$ and gives the expected number of arrivals per unit of time in stationary mode. The variance v of intervals between customer arrivals is calculated as $v = 2\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - \lambda^{-2}$, the squared coefficient c_{var} of variation is equal to $2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$, while the coefficient c_{cor} of correlation of successive intervals between arrivals is given by

$$c_{\text{cor}} = \frac{1}{v}(\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}D_1(-D_0)^{-1}\mathbf{e} - \lambda^{-2}).$$

For more information about the MAP, its special cases, properties and related research, see the works of Lucatoni (1991) and Chakravarthy (2001). The usefulness of the MAP in modeling customer flows in telecommunication systems is mentioned by Heyman and Lucantoni (2003), as well as Klemm *et al.* (2003).

It is assumed that basically customers have to receive service in batches of size N, where N is a certain integer fixed in advance. Below we assume by default that $N \geq 2$. However, results for $N = 1$ are easily obtained from the given formulas as well. Note that $N = 1$ corresponds to the usual service of customers one by one. Due to batch service, an arriving customer has a chance to start service immediately upon arrival only if it arrives when the server is idle and there are $N - 1$ customers in the queue. The arrival of such a customer triggers the start of service of a batch containing N customers. If the customer arrives when the server is busy or idle and the number of customers in the queue is less than $N - 1$, then the customer is placed in the buffer. The capacity of the buffer is infinite. Customers in the buffer are placed in the order of their arrival. The discipline of selecting customers from the buffer at the service completion moment is defined as follows. If at this moment at least N customers are staying in the buffer, the batch consisting of exactly N customers starts service. We call such a batch a *block*. If the number of customers in the buffer at the service completion moment is less than N, we call the set of such customers a *pool*. At this moment, the so-called admission period starts. The server resumes service when the number of customers in the pool reaches the value N

(in this case, customers in the pool form a block and the pool becomes empty) or the admission period expires. In the latter case, if the queue is not empty all customers from the pool start service simultaneously. If the queue is empty, a new admission period starts. Therefore, the server can simultaneously provide service to a batch of N customers (block) if the block was present in the buffer at the service completion epoch or is accumulated there during the admission period, or to the batch of n customers, $n = \overline{1, N-1}$, if the admission period expired when the number of customers in the pool was equal to n.

The structure of the queueing system under study is presented in Fig. 1.

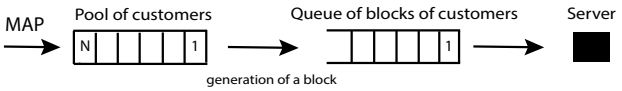

Fig. 1. Structure of the queueing system.

Duration of the admission period has a PH type distribution with an irreducible representation $(\boldsymbol{\tau}, T)$. This means that it is governed by the underlying process $\eta_t^{(0)},\ t \geq 0$, which is a continuous time Markov chain with the state space $\{1, \ldots, M^{(0)}, M^{(0)}+1\}$. The initial state of the process $\eta_t^{(0)},\ t \geq 0$, at the epoch of starting the admission period is determined in the set $\{1, \ldots, M^{(0)}\}$ of transient states by the probabilistic row-vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{M^{(0)}})$. The transitions of the process $\eta_t^{(0)},\ t \geq 0$, within the set $\{1, \ldots, M^{(0)}\}$ do not lead to admission period completion and their intensities are defined by the sub-generator T of size $M^{(0)} \times M^{(0)}$. The intensities of transition to the absorbing state $M^{(0)}+1$, which lead to admission period completion, are defined by the vector $\mathbf{T}_0 = -T\mathbf{e}$. The admission period time distribution function has the form $T(x) = 1 - \boldsymbol{\tau} e^{Tx}\mathbf{e}$. The Laplace–Stieltjes transform $\int_0^\infty e^{-sx}\,\mathrm{d}T(x)$ of this distribution function is $\boldsymbol{\tau}(sI - T)^{-1}\mathbf{T}_0$. The average length of the admission period is given by

$$r_1 = \boldsymbol{\tau}(-T)^{-1}\mathbf{e}.$$

We will further call the value $\mu = r_1^{-1}$ the intensity of admission. The matrix $T + \mathbf{T}_0\boldsymbol{\tau}$ is assumed to be irreducible. A more detailed description of the PH type distribution and its partial cases can be found, e.g., in the book of Neuts (1981).

Duration of the simultaneous service of $n,\ n = \overline{1, N}$, customers has a PH type distribution with an irreducible representation $(\boldsymbol{\beta}^{(n)}, S^{(n)})$. The underlying process of this distribution is $\eta_t^{(n)},\ t \geq 0$, with the finite state space of transient states $\{1, \ldots, M^{(n)}\}$. The average service time of a group of n customers is defined by

$$b_1^{(n)} = \boldsymbol{\beta}^{(n)}(-S^{(n)})^{-1}\mathbf{e}, \quad n = \overline{1, N}.$$

The problem of fitting the measurements of arrival and service processes in real world systems with a Markovian arrival process and a PH distribution can be solved by analogy with the works of Casale *et al*. (2010) and Mèszáros *et al*. (2014). The aim of our further analysis is to evaluate the impact of the value of threshold N and intensity of admission μ on system performance.

3. Process of system states

It can be seen that the dynamics of the system under study are completely described by the multi-dimensional process

$$\xi_t = \{i_t, m_t, r_t, \eta_t^{(r_t)}, \nu_t\}, \quad t \geq 0,$$

where

- i_t is the number of customer batches in the system, $i_t \geq 0$; the number i_t includes one batch in service, if any, and $i_t - 1$ blocks in the queue, if any;
- m_t is the number of customers in the pool, $m_t = \overline{0, N-1}$;
- r_t is the number of customers in service: $r_t = 0$ if $i_t = 0$ and, consequently, the admission period is in progress, and $r_t = \overline{1, N}$ if $i_t \geq 1$;
- η_t is the state of the underlying process of the PH process of customer admission, $\eta_t^{(0)} = \overline{1, M^{(0)}}$, or the state of the underlying process of the PH process of customer service, $\eta_t^{(r_t)} = \overline{1, M^{(r_t)}}$, $r_t = \overline{1, N}$;
- ν_t is the state of the underlying process of the MAP, $\nu_t = \overline{0, W}$.

Note that the total number of customers in the system at an arbitrary moment t is equal to $r_t + (i_t - 1)N + m_t$ if $i_t \geq 1$ or to m_t if $i_t = 0$.

Given all the above assumptions, the five-dimensional process ξ_t is an irreducible continuous time Markov chain with one component (i_t) having the infinite state space and four finite components. Its state space is defined by

$$\begin{aligned}&\left\{\left(0, m, 0, \eta^{(0)}, \nu\right), \eta^{(0)} = \overline{1, M^{(0)}}\right\}\\ \cup &\left\{\left(i, m, r, \eta^{(r)}, \nu\right), i \geq 1, r = \overline{1, N}, \eta^{(r)} = \overline{1, M^{(r)}}\right\},\\ &m = \overline{0, N-1}, \quad \nu = \overline{0, W}.\end{aligned}$$

To analyse the behavior and properties of the Markov chain ξ_t, we have to compute the infinitesimal generator of the chain. Let us denote this generator by $\mathbf{Q}$. The diagonal entries $\mathbf{Q}_{(i,m,r,\eta,\nu),(i,m,r,\eta,\nu)}$ are negative. The modulus of each diagonal entry defines the intensity of departure of the Markov chain from the corresponding

state of the Markov chain. The non-diagonal entry $\mathbf{Q}_{(i,m,r,\eta,\nu),(i',m',r',\eta',\nu')}$ is non-negative and defines the intensity of transition of the Markov chain from the state (i,m,r,η,ν) to the state (i',m',r',η',ν').

To simplify the structure of generator $\mathbf{Q}$ and follow the traditional methodology of the analysis of multi-dimensional Markov chains, it is convenient to make lexicographic enumeration of the states of the Markov chain ξ_t and compose all the states of the chain having value (i,m,r) of the first three components to the *macro-states* $(0,m,0), m=\overline{0,N-1}$, and $(i,m,r), m=\overline{0,N-1}, r=\overline{1,N}$. The macro-state $(0,m,0),\ m=\overline{0,N-1}$, contains

$$K^{(0)}=NM^{(0)}\left(W+1\right)$$

states and the macro-state $(i,m,r), m=\overline{0,N-1}, r=\overline{1,N}$, consists of

$$K^{(r)}=NM^{(r)}\left(W+1\right)$$

states. Analogously, we will compose the macro-states (i,m,r) to *extra-states* $(0,m)\equiv(0,m,0)$, $(i,m)\equiv((i,m,1),\ldots,(i,m,N)), i\geq 1$, and then we will form *super-states* 0 as a composition of extra-states $(0,m),\ m=\overline{0,N-1}$, and i as a composition of extra-states $(i,m),\ m=\overline{0,N-1},\ i\geq 1$.

Lemma 1. *The generator $\mathbf{Q}=(\mathbf{Q}_{i,j})$, where $i\geq 0,\ \max\{0,i-1\}\leq j\leq i+1$, has a three-block diagonal structure:*

$$\mathbf{Q}=\begin{pmatrix}\mathbf{Q}_{0,0} & \mathbf{Q}_{0,1} & O & O & \ldots\\ \mathbf{Q}_{1,0} & \mathbf{Q}_{1,1} & \mathbf{Q}_{1,2} & O & \ldots\\ O & \mathbf{Q}_{2,1} & \mathbf{Q}_{2,2} & \mathbf{Q}_{2,3} & \ldots\\ \vdots & \vdots & \vdots & \vdots & \ddots\end{pmatrix},$$

where non-zero blocks $\mathbf{Q}_{i,j}$ defining the intensities of transitions from super-state i to super-state j, $j=\max\{0,i-1\},i,i+1$, are defined as follows:

- $\mathbf{Q}_{0,0}$ *is a two-block diagonal matrix defined by*

$$\mathbf{Q}_{0,0}=I_N\otimes(\mathbf{T}\oplus D_0)+\widehat{I}_N\otimes(\mathbf{T}_0\tau\otimes I_{\overline{W}})+\mathbf{E}_N^{+}\otimes(I_{M^{(0)}}\otimes D_1),$$

where I_k is the identity matrix of order k, $\otimes$ is the symbol of the Kronecker product of matrices, $\oplus$ is the symbol of the Kronecker sum of matrices, $\widehat{I}_k=\operatorname{diag}\{1,0,\ldots,0\}$, where $\operatorname{diag}\{A_k,\ k=\overline{1,K}\}$, is the diagonal matrix with diagonal entries listed in brackets, $\mathbf{E}_N^{+}=\operatorname{diag}^{+}\{1,\ldots,1\}$ where $\operatorname{diag}^{+}\{A_k,\ k=\overline{1,K}\}$ is the square matrix with the entries above the main diagonal listed in brackets and all other entries equal to 0;

- $\mathbf{Q}_{0,1}$ *is a block matrix having a structure presented below:*

$$\mathbf{Q}_{0,1}=\begin{pmatrix}O & O & \cdots & O\\ (\mathbf{Q}_{0,1})_{1,0} & & & \\ (\mathbf{Q}_{0,1})_{2,0} & \vdots & \ddots & \vdots\\ \vdots & & & \\ (\mathbf{Q}_{0,1})_{N-1,0} & O & \cdots & O\end{pmatrix},$$

where

$$(\mathbf{Q}_{0,1})_{m,0}=\left(\underbrace{O,\ldots,O}_{m-1},\mathbf{T}_0\otimes\boldsymbol{\beta}^{(m)}\otimes I_{\overline{W}},\underbrace{O,\ldots,O}_{N-m}\right),\quad m=\overline{1,N-2}$$

and

$$(\mathbf{Q}_{0,1})_{N-1,0}=\Big(\underbrace{O,\ldots,O}_{N-2},\mathbf{T}_0\otimes\boldsymbol{\beta}^{(N-1)}\otimes I_{\overline{W}},\ \mathbf{e}_{M^{(0)}}\otimes\boldsymbol{\beta}^{(N)}\otimes D_1\Big);$$

- $\mathbf{Q}_{i,i}, i\geq 1$ *is a two-block diagonal matrix with the diagonal blocks defined by*

$$(\mathbf{Q}_{i,i})_{m,m}=\operatorname{diag}\left\{\mathbf{S}^{(r)}\oplus D_0,\ r=\overline{1,N}\right\},\quad m=\overline{0,N-1},$$

and the blocks above the main diagonal defined by

$$(\mathbf{Q}_{i,i})_{m,m+1}=\operatorname{diag}\left\{I_{M^{(r)}}\otimes D_1,\ r=\overline{1,N}\right\},\quad m=\overline{0,N-2};$$

- $\mathbf{Q}_{i,i+1}$ *is the matrix defined by*

$$\mathbf{Q}_{i,i+1}=\begin{pmatrix}O & O & \ldots & \ldots & O\\ \vdots & \vdots & \ddots & & \vdots\\ O & \vdots & & \ddots & \vdots\\ (\mathbf{Q}_{i,i+1})_{N-1,0} & O & \ldots & \ldots & O\end{pmatrix},$$

where $(\mathbf{Q}_{i,i+1})_{N-1,0}$ is the matrix of the form

$$(\mathbf{Q}_{i,i+1})_{N-1,0}=\operatorname{diag}\left\{I_{M^{(r)}}\otimes D_1,\ r=\overline{1,N}\right\};$$

- $\mathbf{Q}_{i,i-1}$ *is the matrix defined by*

$$\mathbf{Q}_{i,i-1}=\operatorname{diag}\{(\mathbf{Q}_{i,i-1})_{m,m},m=\overline{0,N-1}\},$$

with

$$(\mathbf{Q}_{i,i-1})_{m,m} = \begin{pmatrix} O & \dots & O & \mathbf{S}_0^{(1)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \end{pmatrix};$$

- $\mathbf{Q}_{1,0}$ *is a square matrix having* N *block rows and block columns of the form*

$$\mathbf{Q}_{1,0} = \operatorname{diag}\{(\mathbf{Q}_{1,0})_{m,m}, m = \overline{0, N-1}\},$$

with

$$(\mathbf{Q}_{1,0})_{m,m} = \begin{pmatrix} \mathbf{S}_0^{(1)} \otimes \boldsymbol{\tau} \otimes I_{\overline{W}} \\ \vdots \\ \mathbf{S}_0^{(N)} \otimes \boldsymbol{\tau} \otimes I_{\overline{W}} \end{pmatrix}.$$

It is easy to see that for $i \geq 1$ *the expressions of the blocks* $\mathbf{Q}_{i,i}$, $\mathbf{Q}_{i,i-1}$ *and* $\mathbf{Q}_{i,i+1}$ *do not depend on* i*, which means that the Markov chain* ξ_t *belongs to the well-known class of quasi-birth-and-death processes, (see Neuts, 1981).*

Write $\mathbf{Q}_{i,i} = \mathbf{Q}^0$, $\mathbf{Q}_{i,i-1} = \mathbf{Q}^-$ *and* $\mathbf{Q}_{i,i+1} = \mathbf{Q}^+$. *The structure of generator is the following:*

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{0,0} & \mathbf{Q}_{0,1} & O & O & \dots \\ \mathbf{Q}_{1,0} & \mathbf{Q}^0 & \mathbf{Q}^+ & O & \dots \\ O & \mathbf{Q}^- & \mathbf{Q}^0 & \mathbf{Q}^+ & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

4. Ergodicity condition

The ergodicity condition is stated in the following theorem.

Theorem 1. *The Markov chain* ξ_t *is ergodic if the inequality*

$$\lambda b_1^{(N)} < N \tag{1}$$

is fulfilled, and it is non-ergodic if

$$\lambda b_1^{(N)} > N.$$

Here λ *is the fundamental rate of the MAP and* $b_1^{(N)} = \boldsymbol{\beta}^{(N)}(-S^{(N)})^{-1}\mathbf{e}$ *is the average service duration of a batch consisting of* N *customers.*

Proof. From the work of Neuts (1981) it follows that the criterion of the ergodicity of the Markov chain ξ_t is the fulfillment of the inequality

$$\mathbf{y}Q^-\mathbf{e} > \mathbf{y}Q^+\mathbf{e}, \tag{2}$$

where the vector $\mathbf{y}$ is the unique solution of the system of linear algebraic equations

$$\mathbf{y}\left(Q^- + Q^0 + Q^+\right) = \mathbf{0}, \quad \mathbf{y}\mathbf{e} = 1. \tag{3}$$

It is easy to check that the matrix

$$\mathbf{V} = Q^- + Q^0 + Q^+$$

has the following structure:

$$\mathbf{V} = \begin{pmatrix} A & A' & & \\ & \ddots & \ddots & \\ & & \ddots & A' \\ A' & & & A \end{pmatrix},$$

where

$$A = \begin{pmatrix} S^{(1)} \oplus D_0 & & \mathbf{S}_0^{(1)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \\ & \ddots & \vdots \\ & & \ddots \mathbf{S}_0^{(N-1)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \\ & & S^{(N)} \oplus D_0 + \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}} \end{pmatrix},$$

$$A' = \begin{pmatrix} I_{M^{(1)}} \otimes D_1 & & \\ & \ddots & \\ & & I_{M^{(N)}} \otimes D_1 \end{pmatrix}.$$

Let us find a solution to the system (3) rewritten in the form

$$\mathbf{y}\mathbf{V} = \mathbf{0}, \quad \mathbf{y}\mathbf{e} = 1. \tag{4}$$

It is clear that the vector $\mathbf{y}$ has the structure $\mathbf{y} = (\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{N-1})$, where $\mathbf{y}_m = (\mathbf{y}_m^{(1)}, \dots, \mathbf{y}_m^{(N)})$, $m = \overline{0, N-1}$. By direct substitution of this form of the vector $\mathbf{y}$ to the system (4), it can be verified that the vectors $\mathbf{y}_m$, $m = \overline{0, N-1}$, have the following form:

$$\mathbf{y}_m = (\mathbf{0}, \dots, \mathbf{0}, \mathbf{y}_m^{(N)}),$$

where the vectors $\mathbf{y}_m^{(N)}$, $m = \overline{0, N-1}$, satisfy the system of equations

$$\mathbf{y}_{m+1}^{(N)}\left[S^{(N)} \oplus D_0 + \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}}\right] + \mathbf{y}_m^{(N)}(I_{M^{(N)}} \otimes D_1) = \mathbf{0}, \quad m = \overline{0, N-2},$$

$$\mathbf{y}_0^{(N)}\left[S^{(N)} \oplus D_0 + \mathbf{S}_0^{(N)} \otimes \boldsymbol{\beta}^{(N)} \otimes I_{\overline{W}}\right] + \mathbf{y}_{N-1}^{(N)}(I_{M^{(N)}} \otimes D_1) = \mathbf{0},$$

$$\sum_{m=0}^{N-1} \mathbf{y}_m^{(N)} \mathbf{e} = 1.$$

Again by direct substitution, it is possible to verify that the solution of this system of equations is the following:

$$\mathbf{y}_m^{(N)} = \frac{(\boldsymbol{\beta}^{(N)}(-S^{(N)})^{-1}) \otimes \boldsymbol{\theta}}{N b_1^{(N)}}, \quad m = \overline{0, N-1}. \quad (5)$$

After substitution of (5) into inequality (2) and some algebraic manipulations, we get inequality (1). The theorem is proven. ■

Remark 1. The ergodicity (stability) condition for any queueing system is defined by its ability to reduce the number of customers in the system in a situation when this number is huge (the system is overloaded). For the system under study, when it is overloaded, the average number of customers arriving during the service time is equal to $\lambda b_1^{(N)}$ (here $b_1^{(N)}$ is the average service time of a batch of exactly N customers), while the number of customers departing from the system at the service completion moment is given by N. Thus, an intuitively clear condition of system ergodicity should be of the form $\lambda b_1^{(N)} < N$ which coincides with strictly proven condition (1).

The *throughput* of the system (the maximal intensity of the customer flow that can be successfully processed by the system), which is one of the main performance measures of the system, is equal to $N/b_1^{(N)}$.

5. Computation of the stationary distribution of system states and expressions for key performance indices of the system

Further we assume that the inequality (1) is fulfilled. Then a stationary distribution of the Markov chain ξ_t exists. Denote the stationary state probabilities of the chain as

$$\boldsymbol{\pi}(i, m, r, \eta, \nu) = \lim_{t \to \infty} P\{i_t = i,\ m_t = m,\ r_t = r,\ \eta_t = \eta,\ \nu_t = \nu,\},$$
$$i \geq 0, \quad m = \overline{0, N-1}, \quad \nu = \overline{0, W},$$

with $\eta = \overline{1, M^{(0)}}$ if $r = 0$, and $\eta = \overline{1, M^{(r)}}$ if $r = \overline{1, N}$.

Let $\boldsymbol{\pi}(i, m, r)$ be the row vector of the probabilities of states belonging to the macro-state (i, m, r), $\boldsymbol{\pi}(i, m)$ be the row vector of the probabilities of states belonging to the extra-state (i, m), and $\boldsymbol{\pi}_i$ be the row vector of the probabilities of states belonging to the super-state $i,\ i \geq 0$.

Theorem 2. *The stationary probability vectors $\boldsymbol{\pi}_i$ can be computed as follows:*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_1 \boldsymbol{R}^{i-1}, \quad i \geq 2,$$

where $\boldsymbol{R}$ is a solution to the matrix equation

$$\mathbf{Q}^+ + \boldsymbol{R}\mathbf{Q}^0 + \boldsymbol{R}^2\mathbf{Q}^- = \boldsymbol{O}$$

having the spectral radius strictly less than 1, and the vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ are defined as a solution of the system

$$\boldsymbol{\pi}_0 \mathbf{Q}_{0,0} + \boldsymbol{\pi}_1 \mathbf{Q}_{1,0} = \mathbf{0},$$
$$\boldsymbol{\pi}_0 \mathbf{Q}_{0,1} + \boldsymbol{\pi}_1 (\mathbf{Q}^0 + \boldsymbol{R}\mathbf{Q}^-) = \mathbf{0},$$

subject to the normalizing condition

$$\boldsymbol{\pi}_0 \mathbf{e} + \boldsymbol{\pi}_1 (\mathbf{I} - \boldsymbol{R})^{-1} \mathbf{e} = \mathbf{1}.$$

The proof follows easily from the work of Neuts (1981).

Once all the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, have been computed, we are able to calculate various performance measures of the system:

- the average number of blocks of customers in the system, including the one in service,

$$L = \sum_{i=1}^{\infty} i \boldsymbol{\pi}_i \mathbf{e},$$

- the average number of blocks of customers in the system, excluding the one in service,

$$\tilde{L} = \sum_{i=1}^{\infty} (i-1) \boldsymbol{\pi}_i \mathbf{e} = L - 1 + \boldsymbol{\pi}_0 \mathbf{e},$$

- the average number of customers in the pool at an arbitrary moment,

$$N^{(\text{pool})} = \sum_{i=0}^{\infty} \sum_{m=1}^{N-1} m \boldsymbol{\pi}(i, m) \mathbf{e},$$

- the average number of customers in service at an arbitrary moment,

$$N^{(\text{serv})} = \sum_{r=1}^{N} r \left[\boldsymbol{\pi}_1 (I - R)^{-1} \right]^{(r)} \mathbf{e},$$

where we use the notation

$$[\boldsymbol{\pi}_i]^{(r)} = \sum_{m=0}^{N-1} \boldsymbol{\pi}(i, m, r),$$

- the average number of customers in the system at an arbitrary moment,

$$N^{(\text{syst})} = \sum_{i=1}^{\infty}\sum_{m=0}^{N-1}\sum_{r=1}^{N}((i-1)N+m+r)\boldsymbol{\pi}(i,m,r)\mathbf{e} + \sum_{m=0}^{N-1} m\boldsymbol{\pi}(0,m)\mathbf{e} = N\tilde{L} + N^{(\text{serv})} + N^{(\text{pool})},$$

- the probability that an arbitrary customer immediately starts service upon arrival,

$$P_{\text{imm}} = \lambda^{-1}\boldsymbol{\pi}(0,N-1)(\mathbf{e}_{M_0}\otimes D_1\mathbf{e}_{\overline{W}}),$$

- the probability that the server is idle at an arbitrary moment,

$$P_0 = \boldsymbol{\pi}_0(\mathbf{e}_{M_0}\otimes\mathbf{e}_{\overline{W}}),$$

- the probability that the server is idle at the arbitrary arrival moment,

$$P_0^{(\text{arrival})} = \lambda^{-1}\boldsymbol{\pi}_0(\mathbf{e}_{M_0}\otimes D_1\mathbf{e}_{\overline{W}}).$$

6. Waiting time distribution

In this section, we derive the Laplace–Stieltjes transform (LST) of the waiting time distribution. The waiting time is the time interval from the moment of the arrival of an arbitrary customer to the system until the moment when this customer enters service. To get the LST of the stationary waiting time distribution, we use the method of catastrophes (also known as the method of additional events) (Kesten and Runnenburg, 1956; van Dantzig, 1955). This is a powerful method for the derivation of the LST of distributions of quantities such as waiting times, sojourn times, and a busy period. A catastrophe does not have any physical meaning or any impact on the behavior of the queueing system that is being analyzed.

The notion of the catastrophe in the context of an LST of a distribution is frequently employed due to its nice probabilistic interpretation, which is briefly explained. It is assumed that, independently of the queueing system under study, there is a stream of catastrophes that arrive according to a Poisson process with the rate, say, s. We assume s to be a positive real. It is very easy to extend this to complex s, having the real part that is positive. Suppose that ξ is a continuous random variable with distribution function $F_\xi(t)$. Then, it is obvious that the LST

$$\varphi(s) = \int_0^{\infty} e^{-st}\,\mathrm{d}F_\xi(t)$$

of ξ gives the probability that a catastrophe from the stationary Poisson process with the rate s will not arrive during the time given by ξ. The use of this probabilistic interpretation of an LST of a distribution greatly simplifies obtaining an expression for the LST of the waiting time distribution under study, and below we use this approach.

Let us tag an arbitrary arriving customer and monitor its waiting in a queue. Let $w(s)$ be the LST of the distribution of its waiting time or, in other words, the probability that a catastrophe from the stationary Poisson process with the rate s will not arrive during the waiting time. In order to derive an expression for $w(s)$, we need to introduce the following auxiliary denotations. Let $\boldsymbol{\beta}(0,m,s)$ be the column vector consisting of the LSTs of the time until the tagged customer starts service, conditional that currently the server does not provide service (i.e., the admission period is in progress), m customers stay in the pool and the underlying processes of the admission period and arrivals have the corresponding states. Let $\boldsymbol{\beta}(i,m,r,s)$ be the column vector consisting of the LSTs of the time until the tagged customer starts service, conditional that currently the server provides service, there are i blocks in the system, m customers stay in the pool, the current service is provided to the batch consisting of r customers and the underlying processes of service and arrivals have the corresponding states.

Recursive formulas for the $LSTs$ $\boldsymbol{\beta}(0,m,s)$ and $\boldsymbol{\beta}(i,m,r,s)$ are given in the next two lemmas.

Lemma 2. *The $LSTs$ $\boldsymbol{\beta}(0,m,s)$, $m=\overline{1,N-1}$, are computed from the following backward recursion:*

$$\boldsymbol{\beta}(0,N-1,s) = (sI-T\oplus D_0)^{-1}(T_0\otimes\mathbf{e}_{\overline{W}} + \mathbf{e}_{M_0}\otimes D_1\mathbf{e}_{\overline{W}}),$$

$$\boldsymbol{\beta}(0,m,s) = (sI-T\oplus D_0)^{-1}(T_0\otimes\mathbf{e}_{\overline{W}} + (I_{M_0}\otimes D_1)\boldsymbol{\beta}(0,m+1,s)), \quad m=N-2,N-3,\ldots,1.$$

Proof. The formula

$$\boldsymbol{\beta}(0,m,s) = \int_0^{+\infty} e^{(-sI+(T\oplus D_0))t}\,\mathrm{d}t \times\{T_0\otimes\mathbf{e}_{\overline{W}} + I_{M_0}\otimes D_1\boldsymbol{\beta}(0,m+1,s)\} \quad (6)$$

is obvious from the following reasoning. After the arrival of the tagged customer, which joins the pool and becomes the m-th customer in the pool, during some time t, $0<t<\infty$, a catastrophe does not arrive (the probability of this event is e^{-st}); possible transitions of the underlying process of the admission period do not lead to completion

of this period and their probabilities are given by the matrix e^{Tt}; possible transitions of the underlying process of arrivals do not lead to a new customer arrival and their probabilities are obtained by the matrix $e^{D_0 t}$. The joint probability of the described events is equal to

$$e^{-st}e^{Tt} \otimes e^{D_0 t} = e^{(-sI+(T\oplus D_0))t}.$$

After the moment t, during the interval $(t, t+\mathrm{d}t)$ of infinitesimal length, one of two events may happen: (i) the admission period expires (the probabilities of this event under the fixed states of the underlying process of the admission period are given by the vector $T_0 \otimes \mathbf{e}_{\overline{W}}\,\mathrm{d}t$) and service of the tagged customer starts, therefore the probability that a catastrophe does not arrive during the rest of the waiting time is equal to 1; (ii) a new customer arrives (the probabilities of this event under the fixed states of the underlying process of arrivals are given by the matrix $I_{M_0} \otimes D_1\,\mathrm{d}t$). If $m < N-1$, this customer joins the pool and the probabilities that a catastrophe will not arrive during the rest of the waiting time of the tagged customer are given by the vector $\boldsymbol{\beta}(0, m+1, s)$. If $m = N-1$, the batch consisting of N customers (including the tagged one) starts service. Integrating over t, we get (6). The statement of Lemma 2 stems from (6) noting that

$$\int_0^{+\infty} e^{(-sI+(T\oplus D_0))t}\,\mathrm{d}t = (sI-(T\oplus D_0))^{-1}.$$

■

Lemma 3. *The LSTs $\boldsymbol{\beta}(i,m,r,s)$, $r = \overline{1,N}$, are sequentially computed from the equations*

$$\begin{aligned}&\boldsymbol{\beta}(1,m,r,s)\\ &\quad = C_r(s)\boldsymbol{\beta}(0,m,s)\\ &\qquad + B_r(s)\boldsymbol{\beta}(1,m+1,r,s), \quad m=\overline{1,N-2},\end{aligned}$$

$$\begin{aligned}&\boldsymbol{\beta}(1,N-1,r,s)\\ &\quad = C_r(s)\boldsymbol{\beta}(0,N-1,s)\\ &\qquad + B_r(s)H_r(s)\left(\beta^{(N)}\left(sI-S^{(N)}\right)^{-1}\mathbf{S}_0^{(N)}\right)^{i-1},\end{aligned}$$

and

$$\begin{aligned}&\boldsymbol{\beta}(i,m,r,s)\\ &\quad = A_r(s)\boldsymbol{\beta}(i-1,m,N,s) + B_r(s)\boldsymbol{\beta}(i,m+1,r,s),\\ &\qquad\qquad i>1, \quad m=\overline{1,N-2},\end{aligned}$$

$$\begin{aligned}&\boldsymbol{\beta}(i,N-1,r,s)\\ &\quad = A_r(s)\boldsymbol{\beta}(i-1,N-1,N,s)\\ &\qquad + B_r(s)H_r(s)\left(\beta^{(N)}\left(sI-S^{(N)}\right)^{-1}\mathbf{S}_0^{(N)}\right)^{i-1},\\ &\qquad\qquad i>1,\end{aligned}$$

where

$$A_r(s) = \left(sI - S^{(r)} \oplus D_0\right)^{-1}\left(\left(\mathbf{S}_0^{(r)}\beta^{(N)}\right)\otimes I_{\overline{W}}\right),$$

$$B_r(s) = \left(sI - S^{(r)} \oplus D_0\right)^{-1}\left(I_{M_r}\otimes D_1\right),$$

$$C_r(s) = \left(sI - S^{(r)} \oplus D_0\right)^{-1}\left(\left(\mathbf{S}_0^{(r)}\tau\right)\otimes I_{\overline{W}}\right),$$

$$H_r(s) = \left(\left(sI-S^{(r)}\right)^{-1}\mathbf{S}_0^{(r)}\right)\otimes \mathbf{e}_{\overline{W}}.$$

The proof is analogous to that of Lemma 2.

The next theorem is devoted to formulas for the computation of the LST $w(s)$.

Theorem 3. *The LST $w(s)$ is computed as follows:*

$$\begin{aligned}&w(s)\\ &= P_{imm} + \lambda^{-1}\left[\sum_{i=1}^{\infty}\sum_{r=1}^{N}\pi(i,N-1,r)(I_{M_r}\otimes D_1\mathbf{e})\right.\\ &\quad\times(sI-S^{(r)})^{-1}\boldsymbol{S}_0^{(r)}(\beta^{(N)}(sI-S^{(N)})^{-1}\boldsymbol{S}_0^{(N)})^{i-1}\\ &\quad+\sum_{m=0}^{N-2}\pi(0,m)(I_{M_0}\otimes D_1)\boldsymbol{\beta}(0,m+1,s)\\ &\quad+\sum_{i=1}^{\infty}\sum_{m=0}^{N-2}\sum_{r=1}^{N}\pi(i,m,r)(I_{M_r}\otimes D_1)\\ &\quad\left.\times\boldsymbol{\beta}(1,m+1,r,s)\right].\end{aligned}$$

The proof obviously follows from the law of total probability. The expression $(\beta^{(N)}(sI-S^{(N)})^{-1}\boldsymbol{S}_0^{(N)})$ defines the LST of the service time of a block consisting of N customers, the vector $(sI-S^{(r)})^{-1}\boldsymbol{S}_0^{(r)}$ defines the LST of the residual service time of a batch consisting of r customers. The term P_{imm} accounts for the possibility that the tagged customer starts service immediately upon arrival.

The average waiting time can be easily computed by the formula

$$W_1 = -w'(0).$$

7. Numerical results

In this section, we intend to demonstrate feasibility of the proposed algorithms for computation of steady-state distributions of system states and the waiting time under any fixed set of system parameters; to show the effect of variation in the maximal number N of customers that can be processed simultaneously in a batch; to illustrate the high positive effect of the proposed discipline, which suggests that the idle period of the server may end via accumulation of N customers in the pool or via expiration of a certain random amount of time, whichever occurs first, compared with the discipline standard in

the literature that requires mandatory accumulation of N customers; to demonstrate the necessity to account for the correlation in the arrival process to avoid poor evaluation of system performance.

In our derivations we assumed that the service time of a batch consisting of r customers has a PH distribution with an irreducible representation $(\boldsymbol{\beta}^{(r)}, S^{(r)})$, $r = \overline{1, N}$. To implement the numerical work, now we need to fix concrete dependence of the vectors $\boldsymbol{\beta}^{(r)}$ and sub-generators $S^{(r)}$ on r. Let us assume that the service time of an individual customer has a PH type distribution with an irreducible representation $(\boldsymbol{\beta}, S)$ and the size of the vector $\boldsymbol{\beta}$ is M. Evidently, it is reasonable to set $(\boldsymbol{\beta}^{(1)}, S^{(1)}) = (\boldsymbol{\beta}, S)$. For $r > 1$, depending on the potential real world applications, one may think about many options. For example, the service time of a batch consisting of r customers:

- does not depend on r and is identical, in a stochastic sense, with the service time of an individual customer; this option is quite realistic, e.g., in modeling transportation systems. The travel time of an inter-city bus practically does not depend on the number of passengers in the bus;
- is the sum of r service times of individual customers;
- is the weighted sum of r service times of individual customers, e.g., their average value;
- is defined as the minimum of r service times of individual customers; this may be true in a system where, to guarantee quick delivery of some information, the latter is transmitted simultaneously in r channels;
- is defined as the maximum of r service times of individual customers; this kind of dependence takes place in modern telecommunication networks in some networks, e.g., in multi-rate wireless networks with the protocol IEEE802.11 WLAN.

In our numerical results, we fix the last option. In multi-rate protocols, several mobile stations share the same physical channel. Under the use of such protocols, a group of requests from users can be processed simultaneously in parallel and the processing of the whole group is considered finished if the processing of all individual requests belonging to this group is completed. Therefore, the length of the service period of a group has the distribution of the maximum of several independent random variables, each of which represents the service time of an individual customer belonging to this group. Since the expectation of the maximum of a fixed number of independent random variables is less (and can be much less) than the sum of expectations of these random variables, the average time devoted to service of an arbitrary customer under the proposed service discipline may be much less than such time under the classical service discipline. Thus, the throughput of the system under the proposed service discipline is higher and other performance measures of the system may be much better compared with the classical admission discipline. In our numerical experiment we quantitatively illustrate advantages of multi-rate transmission.

The service time of a batch consisting of r customers is defined as the maximum of r service times of individual customers. Because we assumed that the service time of an individual customer has a PH type distribution with an irreducible representation $(\boldsymbol{\beta}, S)$, we have to compute the distribution of the maximum of r service times having such a distribution. As follows from the work of Dudin *et al.* (2015), this maximum indeed has a PH type distribution and its irreducible representation is recursively computed as follows:

$$\boldsymbol{\beta}^{(n)} = \left(\begin{array}{c|c|c} \boldsymbol{\beta} \otimes \boldsymbol{\beta}^{(n-1)} & \mathbf{0}_{M_{n-1}} & \mathbf{0}_M \end{array} \right),$$

$$S^{(n)} = \left(\begin{array}{c|c|c} S \oplus S^{(n-1)} & \mathbf{S}_0 \otimes I_{M_{n-1}} & I_M \otimes \mathbf{S}_0^{(n-1)} \\ \hline O & S^{(n-1)} & O \\ \hline O & O & S \end{array} \right), \quad n \geq 2,$$

with the initial condition $(\boldsymbol{\beta}^{(1)}, S^{(1)}) = (\boldsymbol{\beta}, S)$, where the dimension M_n of vector $\boldsymbol{\beta}^{(n)}$ is defined by $M_n = (M+1)^n - 1,\ n \geq 1$.

To illustrate the effect of correlation in the arrival process, in our experiments, we will consider first three different MAPs having the same fundamental rate $\lambda = 0.6$ but different coefficients of the correlation of successive inter-arrival times.

The first MAP is a stationary Poisson process. It is defined by

$$D_0 = \mathrm{diag}\{-0.6, -0.6\}, \quad D_1 = \begin{pmatrix} 0 & 0.6 \\ 0.6 & 0 \end{pmatrix}.$$

The coefficient of variation of inter-arrival times is equal to 1. The coefficient of correlation of successive inter-arrival times is equal to zero, so we will code this process as MAP_0.

The second MAP, coded as $\mathrm{MAP}_{0.2}$, has a coefficient of correlation $c_{\mathrm{cor}} = 0.2$ and the squared coefficient of variation 12.34. It is defined by the matrices

$$D_0 = \mathrm{diag}\{-0.81156, -0.026346\},$$
$$D_1 = \begin{pmatrix} 0.80616 & 0.0054 \\ 0.014676 & 0.01167 \end{pmatrix}.$$

The third MAP, coded as $\text{MAP}_{0.38}$, has a coefficient of correlation $c_{\text{cor}} = 0.38$ and coefficient of variation $c^2_{\text{var}} = 12.39$. It is defined by the matrices

$$D_0 = \begin{pmatrix} -2.016 & 0 \\ 0.0006 & -0.0654 \end{pmatrix},$$
$$D_1 = \begin{pmatrix} 1.995 & 0.021 \\ 0.0072 & 0.0576 \end{pmatrix}.$$

As the main performance measure of the system under study in our experiments we consider the average waiting time W_1 of an arbitrary customer, while the other performance measures listed in Section 5 are computed as well. It is worth noting that numerous results of various numerical experiments show that the well known Little formula is valid for the system under study in the following form:

$$W_1 = \lambda^{-1}(N\tilde{L} + N^{(\text{pool})}).$$

In all the experiments, we fix $\lambda = 0.6$ as the fundamental rate of the MAP, μ as the intensity of the exponential distribution of the admission period, $0 < \mu \leq 25$, while the distribution of the service time of an individual customer is Erlangian of order 2 with the mean value equal to 1. We investigate the dependence of W_1 on the intensity μ at varying pool capacity N.

Figure 2 shows dependencies of W_1 on μ for MAP_0 and values of N equal to 2,3,4,5. The value of W_1 for $N = 1$ does not depend on μ and is equal to 1.125.

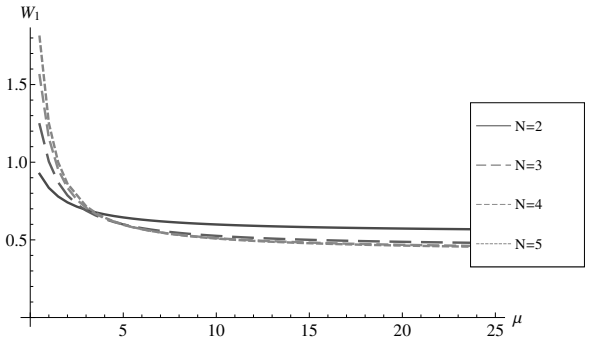

Fig. 2. Average waiting time W_1 for different values of admission rate μ and different dimensions N of the pool when corr $= 0$.

Careful examination of this figure reveals some interesting observations as summarized below:

- If in the classical strategy, which assumes the possibility to start service only when the queue length reaches the level N, we introduce a chance to be served also when a random admission period expires, this essentially decreases the average waiting time. The classical strategy corresponds to the infinite length of the admission period (intensity of the admission period expiration equal to 0). It is evident from Fig. 2, that the increase in μ essentially decreases the average waiting time.

- For small values of μ, small values of pool capacity N are more preferable. However, when μ becomes larger than some value (about 3.5), large values of N become better. Accordingly, a proper choice of μ is desirable for any value of N.

- The difference between the values of W_1 for various N is significant, especially for small values of μ.

Figure 3 reports the behavior of W_1 with respect to μ for $\text{MAP}_{0.2}$. In this case, for $N = 1$ the value of W_1 again does not depend on μ and is equal to 2.852.

We do not present the straight line for $N = 1$ in the figures to avoid suppression of curves.

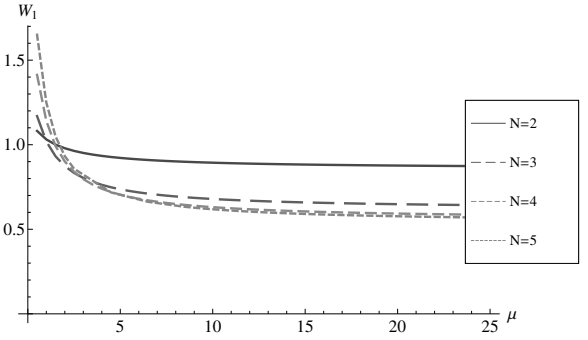

Fig. 3. Average waiting time W_1 for different values of admission rate μ and different dimensions N of the pool when corr $= 0.2$.

From Fig. 3 we can deduce the same conclusions as for Fig. 2. Moreover, we can observe that

(i) The order of the curves for various values of μ can be different and more complicated than the one observed in Fig. 2. Therefore, no "rule of thumb" can be formulated and computation of the average waiting time based on results presented above is mandatory for any available set of N and μ generated by a decision-maker who tries to optimize the system operation.

(ii) The correlation in the arrival process increases the average waiting time.

The latter conclusion becomes much more evident after examining Fig. 4, which depicts W_1 as the function of μ for $\text{MAP}_{0.38}$. In this case, for $N = 1$, the value of W_1 is 77.648.

Figure 4 clearly illustrates the following two facts:

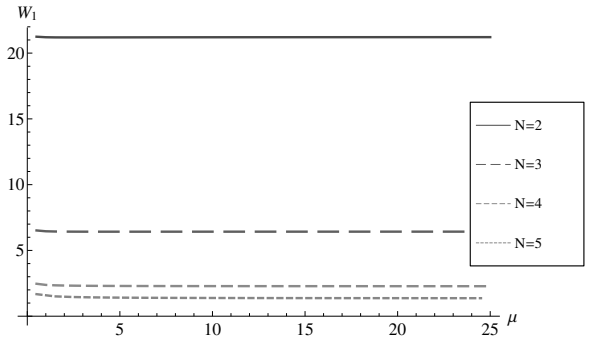

Fig. 4. Average waiting time W_1 for different values of admission rate μ and different dimensions N of the pool when corr $= 0.38$.

(i) Careful account of correlation in an arrival process is vitally important to obtain correct evaluation of system performance measures. Correlation is an important feature of flows in modern telecommunication networks, and it cannot be ignored by assuming that the arrival flow is described by a stationary Poisson arrival process. This ignorance may lead to huge errors.

(ii) The use of batch service can significantly help to improve the quality of system operation. For $N = 1$, we have $W_1 = 77.648$. For $N = 2$, W_1 is about 21, for $N = 3$, W_1 is about 6.4, for $N = 4$, W_1 is about 2.3; for $N = 5$, W_1 is less than 2.

All the three MAPs considered above are artificially constructed to illustrate the effect of correlation and have two states of the underlying process $\nu_t,\ t \geq 0$. Let us repeat the experiment for the MAP obtained as a result of fitting real world traces (see Chydzinski, 2006). The underlying process ν_t of this MAP has five states and is defined by the matrices D_0 and D_1 given by

$$D_0 = \mathrm{diag}\{59620.6, 113826.1, 7892.6, 123563.2, 55428.2\},$$

$$D_1 = \begin{pmatrix} -59793.13 & 38.8 & 30.85 & 0.88 & 102.00 \\ 16.76 & -114709.36 & 97.52 & 398.90 & 370.08 \\ 281.48 & 445.97 & -9487.09 & 410.98 & 456.06 \\ 23.61 & 205.74 & 58.49 & -124162.13 & 311.09 \\ 368.48 & 277.28 & 7.91 & 32.45 & -56114.32 \end{pmatrix}.$$

These matrices are obtained based on information about the generator of the underlying process ν_t and the expression for D_1 as the diagonal matrix presented by Chydzinski (2006). The original matrices are scaled to obtain an MAP having the same fundamental rate $\lambda = 0.6$ as the three MAPs of order 2 which were used to build Figs. 2–4. The coefficients of correlation and variation of the MAP are not changed under the scaling and are as follows: $c_{\mathrm{cor}} = 0.141684$ and $c^2_{\mathrm{var}} = 1.46354$. Thus, the MAP considered of order 5 has the coefficients of correlation and variation intermediate between the values of these coefficients for MAP_0 and $\mathrm{MAP}_{0.2}$. Therefore, one may anticipate that the value of the average waiting time W_1 for various values of N should be intermediate between the values of W_1 for MAP_0 and $\mathrm{MAP}_{0.2}$. However, for $N = 1$ we have that $W_1 = 1.125$ for MAP_0, $W_1 = 2.852$ for $\mathrm{MAP}_{0.2}$, and $W_1 = 6.63691$ for the given MAP of order 5. The high value of W_1 for the MAP of order 5 is easily explained by the existence of two states of the underlying process ν_t, in which the intensity of generation of customers is much higher than in other states. Such irregularity in arrivals implies that sometimes the server is idle but sometimes it loaded. It is important to note that the results of computations for $N = 2, 3, 4, 5$ presented in Fig. 5 show that the negative effect of irregularity in arrivals is essentially mitigated by providing service in groups.

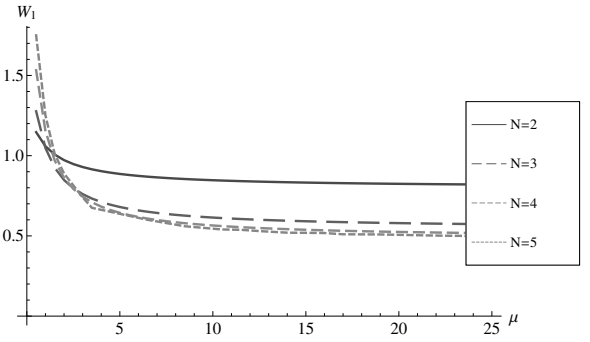

Fig. 5. Average waiting time W_1 for different values of admission rate μ and different dimensions N of the pool for the MAP of order 5.

This confirms the importance of the analysis presented in this paper. Advantages of group service are illustrated and an algorithmic tool is provided for optimal choice of the pair N and μ in situations when the use of large values of N is restricted technically or economically. In telecommunications systems, N can be interpreted as the level of multiplexing or the number of mobile stations, which can share the channel to an access point, and the value of N can be limited by available bandwidth. In applications to transportation systems, N can be interpreted as the capacity of vehicles or minivans which can be leased for passengers delivering. In applications to manufacturing systems, N can be

interpreted as the capacity of pallets used for providing technological operations like heating or cooling some details, etc. After the choice of the appropriate value of the parameter N based on the restrictions on the waiting time of customers and the cost of using the corresponding capacity of bandwidth, our results allow fixing also a suitable value of the parameter μ. The choice of μ is not trivial. If we choose a small value of μ, we benefit from a high coefficient of utilization of the capacities used (bandwidth, vehicles, pallets, etc) but this puts us at risk of providing poor quality of service. A long waiting time can make information to be transmitted outdated, passengers to be delivered to airport miss the flight, details to be processed can lose required properties, etc. But if we choose a large value by μ, we benefit from the usage of advantages of group service and this offers good quality of service, but the coefficient of utilization of the capacities used may be low. One may observe in Figs. 2, 3, 5 that for large values of μ the difference between the values of W_1 for $N = 4$ and $N = 5$ is quite small. Therefore, the presented results can be useful to find some trade-off between the quality of provided service and the provider's expenditures.

8. Conclusion

A novel customer batch service discipline for a single server queue was introduced and analyzed. Service to customers was offered in batches of a certain size N. If the number of customers in the system at the service completion moment is less than N, the server does not start next service until the number of customers in the system reaches this size or the admission period, which limits the idle time of the server, expires, whichever occurs first. The dynamics of such a system are described by a multi-dimensional Markov chain.

Because this chain belongs to the class of QBD, analysis of its steady state distribution is more or less straightforward, while derivation of the Laplace-Stieltjes transform of the waiting time is much more involved. It is implemented with the help of the method of a supplementary event. The dependence of the average waiting time on N, mean value of the admission period and correlation in the arrival process was numerically illustrated. Important conclusions, based on the results of computations, were formulated. The results are going to be extended to the case of batch arrivals similar to the one of Gaidamaka *et al*. (2014), customers impatient during the stay in the pool similar to the case investigated by Dudin *et al*. (2016), a system operating in a random environment similar to the case studied by Kim *et al*. (2014) and a discrete time system similar to the one presented by Atencia (2014).

References

Atencia, I. (2014). A discrete-time system with service control and repairs, *International Journal of Applied Mathematics and Computer Science* **24**(3): 471–484, DOI: 10.2478/amcs-2014-0035.

Bailey, N. (1954). On queueing processes with bulk service, *Journal of the Royal Statistical Society B* **16**(1): 80–87.

Banerjee, A., Gupta, U. and Chakravarthy, S. (2015). Analysis of afinite-buffer bulk-service queue under Markovian arrival process with batch-size-dependent service, *Computers and Operations Research* **60**: 138–149.

Casale, G., Zhang, E. and Smirn, E. (2010). Trace data characterization and fitting for Markov modeling, *Performance Evaluation* **67**(2): 61–79.

Chakravarthy, S. (2001). The batch Markovian arrival process: A review and future work, *in* V.R.E.A. Krishnamoorthy and N. Raju (Eds.), *Advances in Probability Theory and Stochastic Processes*, Notable Publications Inc., Branchburg, NJ, pp. 21–29.

Chydzinski, A. (2006). Transient analysis of the MMPP/G/1/K queue, *Telecommunication Systems* **32**(4): 247–262.

Deb, R. and Serfozo, R. (1973). Optimal control of batch service queues, *Advances in Applied Probability* **5**(2): 340–361.

Downton, F. (1955). Waiting time in bulk service queues, *Journal of the Royal Statistical Society B* **17**(2): 256–261.

Dudin, A., Manzo, R. and Piscopo, R. (2015). Single server retrial queue with adaptive group admission of customers, *Computers and Operations Research* **61**: 89–99.

Dudin, A., Lee, M.H. and Dudin, S. (2016). Optimization of the service strategy in a queueing system with energy harvesting and customers' impatience, *International Journal of Applied Mathematics and Computer Science* **26**(2): 367–378, DOI: 10.1515/amcs-2016-0026.

Gaidamaka, Y., Pechinkin, A., Razumchik, R., Samouylov, K. and Sopin, E. (2014). Analysis of an $M/G/1/R$ queue with batch arrivals and two hysteretic overload control policies, *International Journal of Applied Mathematics and Computer Science* **24**(3): 519–534, DOI: 10.2478/amcs-2014-0038.

Heyman, D. and Lucantoni, D. (2003). Modelling multiple IP traffic streams with rate limits, *IEEE/ACM Transactions on Networking* **11**(6): 948–958.

Kesten, H. and Runnenburg, J. (1956). *Priority in Waiting Line Problems*, Mathematisch Centrum, Amsterdam.

Kim, C., Dudin, A., Dudin, S. and Dudina, O. (2014). Analysis of an $MMAP/PH_1$, $PH_2/N/\infty$ queueing system operating in a random environment, *International Journal of Applied Mathematics and Computer Science* **24**(3): 485–501, DOI: 10.2478/amcs-2014-0036.

Klemm, A., Lindermann, C. and Lohmann, M. (2003). Modelling IP traffic using the batch Markovian arrival process, *Performance Evaluation* **54**(2): 149–173.

Lucatoni, D. (1991). New results on the single server queue with a batch Markovian arrival process, *Communication in Statistics: Stochastic Models* **7**(1): 1–46.

Mèszáros, A., Papp, J. and Telek, M. (2014). Fitting traffic with discrete canonical phase type distribution and Markov arrival processes, *International Journal of Applied Mathematics and Computer Science* **24**(3): 453–470, DOI: 10.2478/amcs-2014-0034.

Neuts, M. (1967). A general class of bulk queues with Poisson input, *The Annals of Mathematical Statistics* **38**(3): 759–770.

Neuts, M. (1981). *Matrix-geometric Solutions in Stochastic Models—An Algorithmic Approach*, Johns Hopkins University Press, Baltimore, MD.

Sasikala, S. and Indhira, K. (2016). Bulk service queueing models—a survey, *International Journal of Pure and Applied Mathematics* **106**(6): 43–56.

van Dantzig, D. (1955). Chaines de markof dans les ensembles abstraits et applications aux processus avec regions absorbantes et au probleme des boucles, *Annales de l'Institut Henri Poincaré* **14**(3): 145–199.

Arianna Brugno was born in 1989. She graduated in mathematics in 2013 from the University of Salerno, Italy. Now she is a PhD student of mathematics there. Her main research area is queueing theory.

Ciro D'Apice is a full professor of mathematical analysis at the Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, Italy. His research interests include variational calculus, homogenization and optimal control, conservation laws and applications to traffic, telecommunication networks, and supply chains, queueing systems and networks, as well as analytical aspects for the temporal and spatial behaviour of solutions of dynamic problems.

Alexander Dudin is the head of the Laboratory of Applied Probabilistic Analysis at Belarusian State University, a professor at the Department Probability Theory and Mathematical Statistics. He is the author of 350 publications including more than 100 papers in top level journals. He is the chairman of the IPC of Belarusian Workshops on *Queueing Theory* and international conferences named after A.F. Terpugov in Siberia. His field of scientific interests includes random processes and queueing theory. He has been invited to the USA, the UK, Germany, France, Holland, Japan, South Korea, India, Russia, China, Sweden and Italy. He received a Belarus Scopus Award in 2013.

Rosanna Manzo is a researcher in mathematical analysis at the Department of Information Engineering, Electrical Engineering and Applied Mathematics of the University of Salerno, Italy. She received her PhD in information engineering from the University of Salerno. Her research areas include fluid-dynamic models for traffic flows on road, telecommunications and supply networks, optimal control and queueing theory.

Received: 19 July 2016
Revised: 6 November 2016
Accepted: 9 November 2016