Int. J. Appl. Math. Comput. Sci., 2017, Vol. 27, No. 2, 367–384
DOI: 10.1515/amcs-2017-0026

amcs

A QUEUEING SYSTEM WITH HETEROGENEOUS IMPATIENT CUSTOMERS AND CONSUMABLE ADDITIONAL ITEMS

JANGHYUN BAEK [a], OLGA DUDINA [b,c], CHESOONG KIM [d,*]

[a]Department of Industrial and Information Systems Engineering
Chonbuk National University, Jeonju, 54896, Republic of Korea
e-mail: jbaek@chonbuk.ac.kr

[b]Research Laboratory of Applied Probabilistic Analysis
Belarusian State University, 4, Nezavisimosti Av., Minsk, 220030, Belarus
e-mail: dudina@bsu.by

[c] RUDN University, 6, Miklukho-Maklaya Str., Moscow, 117198, Russia
e-mail: dudina_olga@email.com

[d]Department of Industrial Engineering
Sangji University, Wonju, Kangwon, 26339, Republic of Korea
e-mail: dowoo@sangji.ac.kr

A single-server queueing system with a marked Markovian arrival process of heterogeneous customers is considered. Type-1 customers have limited preemptive priority over type-2 customers. There is an infinite buffer for type-2 customers and no buffer for type-1 customers. There is also a finite buffer (stock) for consumable additional items (semi-products, half-stocks, etc.) which arrive according to the Markovian arrival process. Service of a customer requires a fixed number of consumable additional items depending on the type of the customer. The service time has a phase-type distribution depending on the type of the customer. Customers in the buffer are impatient and may leave the system without service after an exponentially distributed amount of waiting time. Aiming to minimize the loss probability of type-1 customers and maximize throughput of the system, a threshold strategy of admission to service of type-2 customers is offered. Service of type-2 customer can start only if the server is idle and the number of consumable additional items in the stock exceeds the fixed threshold. Stationary distributions of the system states and the waiting time are computed. In the numerical example, we show some interesting effects and illustrate a possibility of application of the presented results for solution of optimization problems.

Keywords: marked Markovian arrival process, consumable additional items, phase-type distribution, impatient customers.

1. Introduction

Queueing models are useful for solving problems of capacity planning, performance evaluation and optimization of a variety of real life systems, objects and processes. Usually, it is assumed that the service times of successive customers are defined by a sequence of identically distributed random variables that are independent of the arrival process and other entities defining operation of the system. However, sometimes in real life systems there is a dependence of the service process on the presence or absence of some *additional items* (windows, tokens, permissions, connections, threads, sessions, details, semi-finished products, half-stocks, energy, etc.). The standard situation in such systems is as follows. There is a finite stock of additional items. The items are rented by the server from this stock to provide service to a customer. After the service completion, the rented item returns to the stock, i.e., the items are reusable. If the stock is empty at a customer arrival moment, its service is suspended. Such a situation takes place, e.g., in the systems where it is

*Corresponding author

necessary to regulate the speed of customers admission (in telecommunication protocols like sliding window, leaky bucket, etc.) or when service can be provided only with help of some equipment.

Among the related papers, we can mention the paper by Dudin and Klimenok (1996), where the additional item is interpreted as a passive server, and the papers of Kim *et al.* (2009; 2012) where the additional item is interpreted as a token that is necessary to start service of a session in communication network. Similar queueing models arise in the analysis of so called queueing/inventory models, (see Krishnamoorthy *et al.*, 2016c; 2016b; 2016a).

In this paper, we also assume that service of a customer requires the use of additional items, but the items are not reusable, they are consumable. A definite number of items permanently disappear from the stock after beginning service of a customer. If during the service completion epoch the stock does not contain the required number of additional items for the next service, service of customers is suspended. Service is resumed only after the moment when the suitable number of additional items arrives to the system.

We consider the system with two types of customers having different service time distributions and different numbers of additional items required for service. Customers of type-1 have limited preemptive priority over type-2 customers. This means the following. The arrival of a type-1 customer during the epoch when the server provides service to a type-2 customer implies forced termination of service of the type-2 customer and its loss only if the stock contains the required number of additional items for service of this type-1 customer. Otherwise, service of the type-2 customer is not interrupted and the arrived type-1 customer is lost. Besides the evident inconvenience for the type-2 customer, the forced termination of its service leads to the waste of items engaged into service of the terminated type-2 customer. In turn, it may further imply the lack of the items necessary for service of type-1 customers. Thus, the problem of optimal management of admission of low priority customers to the service arises.

In this paper, we consider the following strategy of customers admission. The type-1 customer is always admitted to service if the server is not busy with service of a type-1 customer and there is a sufficient number of items in the stock. Type-2 customers are always admitted to the system. However, a type-2 customer is admitted to service only if all type-2 customers, who arrived earlier than this customer, left the system, the server is idle and the number of items in the stock exceeds some fixed threshold, say, N. Additionally, we assume that type-2 customers are impatient and leave the system (are lost) after a random amount of time. The goal of the paper is the analysis of the dependence of the key performance measures of the system on the value of the threshold. This creates an opportunity to formulate and solve a problem of optimal (with respect to some chosen economical criterion) choice of the threshold N. This problem is not trivial. If N is too small, type-2 customers have easy access to the system, but many type-1 customers, whom we consider as the priority customers probably needed urgent treatment, will be lost due to the lack of required additional items. If N is too large, type-1 customers have easy access to the system, but many type-2 customers will be lost due to the long waiting in the queue and the system will be underutilised.

Examples of possible applications of the model under consideration are as follows:

1. Optimization of operation of a node of a sensor network. The sensor node collects information about some object and generates information units that have to be transmitted to a central node. There are two types of information units. Type-1 corresponds to emergency related information. Type-2 corresponds to routine monitoring of the object parameters. The sensor node has a battery of a small finite capacity and has to harvest energy during its operation from outside (we can mention solar cells, wind turbines, piezo electric cells, radio frequency collectors, etc.). We assume that the energy is slotted to energy units and several energy units are required to transmit one information units. To save energy, especially to have some reserve of energy for transmission of emergency related information, the node sometimes should switch off transmission of information. The time when service is interrupted should be long to earn enough energy units to battery for future transmissions, but it should be pretty short due to obsolescence of information collected by the sensor node. If the waiting time of an information unit exceeds some level, transmission of this information becomes meaningless. The queueing model under study should help to find some trade-off in this situation. Analogous queueing models (but only for homogeneous customers) were formulated, e.g., by Yang and Ulukus (2012b), Sharma *et al.* (2010), Tutuncuoglu and Yener (2012) or Yang and Ulukus (2012a), who gave more information about quite extensive literature concerning the systems with energy harvesting is presented. However, the analysis was provided there in deterministic settings or not in terms of queueing theory. Queueing analysis of a similar system was recently done by Gelenbe (2015). But it is assumed therein that flows of customers and energy units are instantaneously synchronised, so the buffer or (and) the stock should always be empty. In our model, we consider a more a general situation in which the service time is not equal to zero and it is possible that the buffer and the stock may be not empty simultaneously.

2. Design, management and optimization of warehousing and inventory systems. (Manzini *et al.*, 2015) Inventory items arrive at random moments and are stored

in a warehouse of a finite capacity. There are two types of orders arriving to this warehouse. The orders of different types are distinguished by the priority, the number of demanded inventory items and the time required for extracting these items from the warehouse.

3. Performance evaluation of operation of some manufacturing line where some products are produced or repaired. Production of one unit of a product requires several units of semi-finished product. If this semi-finished product is absent at the stock, the production is not possible. In this case, it makes sense to stop a manufacturing line (to provide a vacation) during some period of time and, then, to resume the work after generation of sufficient number of semi-finished products. Vacation may be used for the work related to the manufacturing line maintenance or for power saving.

4. Capacity planning of some medical center where the emerging and elective (scheduled) surgery should be provided. (Cardoen *et al.*, 2010) The server may be interpreted as an operating room or a surgeon or a surgical team. The additional item may be interpreted as a portion of the amount of blood (drug, single-use equipment, etc) necessary for the surgery. In this potential application, a too small value of threshold N may imply refusal of surgery to the critically ill patient and, probably, his/her death. A too large value of threshold N may imply, due to too long waiting, exodus of scheduled patients to competitive medical centers. This leads to the low load of the operating room or the surgical team and a low profit of the medical center gained by providing this type of medical care. Thus, the problem of the optimal choice of the value of threshold N is quite important.

Here we provide analysis of the above described queueing system. We assume that the arrival flows of heterogeneous customers and additional items are stochastic and the service times of customers are random. We do not follow traditional assumptions in the queueing literature that the arrival flows are stationary Poisson and the service time distribution is exponential. We assume that the arrival flows of customers and units are described by a much more general marked Markovian arrival process and a Markovian arrival process, respectively. This allows us to catch possible correlation in the arrival processes and a possibility to have time intervals when customers or (and) additional items arrive rarely or frequently. This is important from the point of view of applications because, e.g., the speed of energy harvesting in a sensor network may essentially vary depending on the sun shine or the speed of the wind.

As will be shown in the numerical results below, correlation in the arrival processes drastically changes the system characteristics compared with those in the system with the stationary Poisson arrival processes having the same mean arrival rate. Concerning the service processes, we assume that the service times of two types of customers have different PH (phase-type) distributions which are much more general than exponential distribution. This is also very important from the point of view of potential applications. The assumption that the service time has the exponential distribution drastically simplifies analysis of the model, but is not realistic in a majority of real world systems. The probability density function of the exponential distribution is maximal at zero. This means that the most probable duration of service time is equal to zero, which is hardly realistic.

Queueing systems similar to our model were considered by Dudin *et al.* (2016) or Zhao and Lian (2011). In the work of Dudin *et al.* (2016), the energy harvesting model was analyzed with one class of customers and the use of exactly one item for service of any customer. Here we consider two classes of customers with different priorities and requirements on the number of items. The model with two classes of customers and the necessity to use items for service provisioning was considered recently in an interesting paper by Zhao and Lian (2011) as a queueing-inventory model. An overview of previous research was presented there. In particular, it is noted there that only models with one class of customers were analysed before, while two classes are assumed by Zhao and Lian (2011).

Three main differences from the model considered in our paper are as follows. We assume that high priority customers are not queued. They have limited preemptive priority over the low priority customers. In the work of Zhao and Lian (2011), both types of customers are queued and high priority customers have non-preemptive priority. We assume that the items arrive at random moments, while the well-known (r, Q) replenishment strategy was assumed by Zhao and Lian (2011). Advantages of that paper are the presentation of so called μb rule for establishing the priorities, the proof of ergodicity condition and the use of matrix analytical methods for computation of the stationary distribution of the system states. In some other aspects, our model is more general than the one by Zhao and Lian (2011): we assume more general arrival processes of customers and items; we assume more general service time distributions; we assume that service of a customer requires not only one, but several items depending on the type of customer; we assume that the low priority customers are impatient. These generalizations essentially complicate the analysis of the model. However, the presented numerical results reveal great importance of these generalizations for adequate modelling of the real world systems.

The rest of the paper consists of the following. In Section 2, the mathematical model is described. In Section 3, the process of the system states is completely defined and the problem of computation of stationary probabilities of the system states is touched. Formulas for

computation of the performance measures of the system are presented in Section 4. The problem of computation of the stationary distribution of the waiting time of an arbitrary type-2 customer is solved in Section 5. Quite interesting numerical illustrations are given and briefly discussed in Section 6. Section 7 concludes the paper.

2. Mathematical model

We consider a single-server queueing system with two types of customers, no buffer for type-1 customers, an infinite buffer for type-2 customers and a buffer (stock) of capacity K for additional items. The structure of the system under study is presented in Fig. 1.

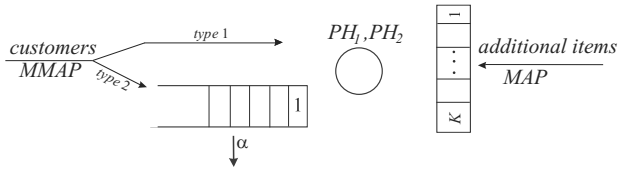


Fig. 1. Queueing system under study.

Customers arrive at the system according to the marked Markovian arrival process (MMAP). The customers in the MMAP are heterogeneous and have different types. The arrival of customers is directed by a stochastic process $\nu_t,\ t \geq 0$, which is an irreducible continuous-time Markov chain with state space $\{0, 1, \ldots, W\}$. The sojourn time of this chain in the state ν is exponentially distributed with a positive finite parameter $\lambda^{(\nu)}$. When the sojourn time in the state ν expires, with probability $p^{(0)}_{\nu,\nu'}$ the process ν_t jumps to the state ν' without generation of a customer, $\nu, \nu' = \overline{0, W}$, $\nu \neq \nu'$, and with probability $p^{(l)}_{\nu,\nu'}$ the process ν_t jumps to the state ν' with a generation of type-l customer, $l = 1, 2$, $\nu, \nu' = \overline{0, W}$. Here and in the sequel the notation like $\nu = \overline{0, W}$ means that the parameter ν takes values from the set $\{0, 1, \ldots, W\}$.

The behavior of the MMAP is completely characterized by the matrices $D_l,\ l = 0, 1, 2$, defined by the entries $(D_l)_{\nu,\nu'} = \lambda^{(\nu)} p^{(l)}_{\nu,\nu'}$, $\nu, \nu' = \overline{0, W}$, $l = 1, 2$, and $(D_0)_{\nu,\nu} = -\lambda^{(\nu)}, \nu = \overline{0, W}$, $(D_0)_{\nu,\nu'} = \lambda^{(\nu)} p^{(0)}_{\nu,\nu'}$, $\nu, \nu' = \overline{0, W}$, $\nu \neq \nu'$. The matrix $D(1) = D_0 + D_1 + D_2$ represents the generator of the process $\nu_t, t \geq 0$.

The average total arrival intensity λ is defined by $\lambda = \boldsymbol{\theta}(D_1 + D_2)\mathbf{e}$, where $\boldsymbol{\theta}$ is the invariant vector of the stationary distribution of the Markov chain $\nu_t, t \geq 0$. The vector $\boldsymbol{\theta}$ is the unique solution to the system $\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$. Here $\mathbf{e}$ denotes a column vector consisting of 1's, and $\mathbf{0}$ is a zero row vector. The average arrival intensity λ_l of type-l customers is defined by $\lambda_l = \boldsymbol{\theta}\, D_l \mathbf{e}$, $l = 1, 2$.

The squared integral (without differentiating the types of customers) coefficient of variation of intervals between successive arrivals is given as $c_{var} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$. The squared coefficient of variation of inter-arrival times of type-l customers is given as $c^{(l)}_{var} = 2\lambda_l \boldsymbol{\theta}(-D_0 - D_1^{(\bar{l})})^{-1}\mathbf{e} - 1,\ \ \bar{l} \neq l, \bar{l}, l = 1, 2$. The integral coefficient of correlation of two successive intervals between arrivals is given as $c_{cor} = (\lambda\boldsymbol{\theta}(-D_0)^{-1}(D(1) - D_0)(-D_0)^{-1}\mathbf{e} - 1)/c_{var}$. More information about the MMAP and related research is given, e.g., by He (1996).

We assume that k_l additional items are required for service of each type-l customer, $l = 1, 2$. Consequently, at the moment when a type-l customer is chosen for service, the number of additional items in the stock decreases by k_l, $l = 1, 2$.

Type-1 customers are assumed to be priority customers and have limited preemptive priority over type-2 customers. An arriving type-1 customer is not accepted for service (is lost) only if the server provides service to another type-1 customer or the number of additional items in the stock is less than k_1. If the server provides service to a type-2 customer at the moment of the arrival of a type-1 customer, service of the type-2 customer is immediately terminated, it leaves the system permanently (is lost) and the additional items engaged into service are lost as well, provided there are at least k_1 additional items available. Different scenarios of the system behavior during an arbitrary type-1 customer arrival epoch are illustrated in Figs. 2–5:

- The server is busy with a type-1 customer, the arriving customer is lost (see Fig. 2).

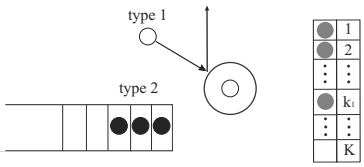


Fig. 2. Arrival of a type-1 customer when the server is busy with a type-1 customer.

- the number of additional items in the stock is less than k_1, the arriving customer is lost (see Fig. 3). If at this moment the server provides service to a type-2 customer, the service is not terminated.

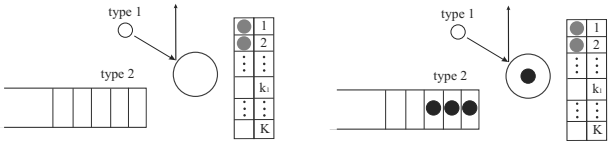


Fig. 3. Arrival of a type-1 customer when the number of additional items in the stock is less than k_1.

- The number of additional items in the stock is greater than or equal to k_1 and the server is free, the arriving customer occupies the server (see Fig. 4). The left part of Fig. 4 illustrates the case $k_1 < k_2$ (therefore, the server can be idle while several type-2 customers are waiting in the buffer) and the right part illustrates the case $k_1 \geq k_2$.

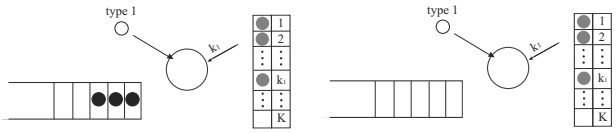


Fig. 4. Arrival of a type-1 customer when the number of additional items in the stock is greater than or equal to k_1 and the server is free.

- The number of additional items in the stock is greater than or equal to k_1 and the server is busy with a type-2 customer, the arriving type-1 customer occupies the server. The type-2 customer, whose service is terminated, is lost (see Fig. 5).

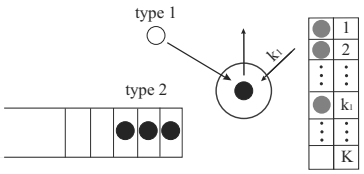


Fig. 5. Arrival of a type-1 customer when the number of additional items in the stock is greater than or equal to k_1 and the server is busy with a type-2 customer.

The arriving type-2 customer can start service only if the server is idle and the number of additional items in the stock is greater than some preassigned threshold N, $N \geq k_2$. Otherwise, this customer joins the buffer. Different scenarios of the system behavior at an arbitrary type-2 customer arrival epoch are illustrated in Figs. 6 and 7:

- The number of additional items in the stock is greater than N, $k_2 - 1 \leq N < K$, and the server is free, the arriving customer occupies the server and the number of additional items decreases by k_2 (see Fig. 6).

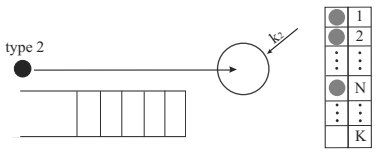


Fig. 6. Arrival of a type-2 customer when the number of additional items in the stock is greater than N and the server is free.

- The number of additional items in the stock is less than or equal to N or the server is busy, the arriving type-2 customer is placed into the buffer (see Fig. 7).

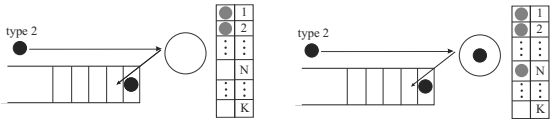


Fig. 7. Arrival of a type-2 customer when the number of additional items in the stock is less than or equal to N or the server is busy.

Type-2 customers are picked up from the queue according to the first in–first out discipline. The type-2 customer from the buffer can start service at an arbitrary service completion moment only if it is the first in the buffer and there are at least N additional items in the stock.

Additional items arrive at the system according to a Markovian arrival process (MAP). Arrivals in the MAP are directed by an irreducible continuous time Markov chain ζ_t, $t \geq 0$, with the finite state space $\{0, 1, \ldots, V\}$. The MAP is defined by the matrices H_0 and H_1. Let us denote as λ_e the average intensity of the MAP. For more information about the MAP, (see, e.g., Chakravarthy, 2001). If at the arrival epoch of additional item, the stock of items is full, then the arriving item is lost.

The service time of an arbitrary type-l customer has a PH distribution with an irreducible representation $(\boldsymbol{\beta}_l, S_l)$, $l = 1, 2$. This service time can be interpreted as a time until the underlying Markov process $\eta_t^{(l)}, t \geq 0$, with a finite state space $\{1, \ldots, M_l, M_l + 1\}$ reaches the single absorbing state $M_l + 1$ conditioned on the fact that the initial state of this process is selected among the transient states $\{1, \ldots, M_l\}$ according to the probabilistic row vector $\boldsymbol{\beta}_l = (\beta_1^{(l)}, \ldots, \beta_{M_l}^{(l)})$. The transition rates of the process $\eta_t^{(l)}$ within the set $\{1, \ldots, M_l\}$ are defined by the sub-generator S_l, and the transition rates into the absorbing state (what leads to service completion) are given by the entries of the column vector $\mathbf{s}_0^{(l)} = -S_l\mathbf{e}$. The mean service time of type-l customer is calculated by $b_1^{(l)} = \boldsymbol{\beta}_l(-S_l)^{-1}\mathbf{e}$, $l = 1, 2$. The squared coefficient of variation is given by $c_{var}^{(l)} = b_2^{(l)}/(b_1^{(l)})^2 - 1$ where $b_2^{(l)} = 2\boldsymbol{\beta}_l(-S_l)^{-2}\mathbf{e}$. For more information about the PH distribution and its usefulness, see, e.g., the work of Neuts (1981).

Type-2 customers in the buffer are impatient, i.e., customers leave the buffer after an exponentially distributed time with the parameter α, $0 < \alpha < \infty$, after arrival due to a lack of service.

For the reader's convenience, we summarize the main notation that characterizes the system in Table 1.

3. Process of the system states

It is easy to see that the behavior of the system under study is described in terms of the following regular irreducible continuous-time Markov chain

$$\xi_t = \{i_t, r_t, k_t, \nu_t, \zeta_t, \eta_t\}, \quad t \geq 0,$$

where, during the epoch $t,\ t \geq 0$,

- i_t is the number of type-2 customers in the buffer, $i_t \geq 0$;
- r_t is an indicator of the status of the server: $r_t = 0$ if the server is free, $r_t = r$ if the server serves type-r customer, $r = 1, 2$;
- k_t is the number of additional items, $k_t = \overline{0, K}$;
- ν_t is the state of the underlying process of the MMAP, $\nu_t = \overline{0, W}$;
- ζ_t is the state of the underlying process of the MAP, $\zeta_t = \overline{0, V}$;
- η_t is the state of the PH service process, with $\eta_t = \overline{1, \delta_{r_t=1} M_1 + \delta_{r_t=2} M_2}$.

Here we write

$$\delta_{[\text{condition}]} = \begin{cases} 1 & \text{if the condition holds true,} \\ 0 & \text{otherwise.} \end{cases}$$

Table 1. Notation.

K	the capacity of buffer (stock) for additional items
D_l, $l = 0, 1, 2$	the square matrices of size $W + 1$ that characterize the MMAP arrival flow of customers
λ_l, $l = 1, 2$	the average arrival intensity of type-l customers
H_l, $l = 0, 1$	the square matrices of size $V + 1$ that characterize the MAP arrival flow of additional items
λ_e	the average arrival intensity of additional items
k_l, $l = 1, 2$	the number of additional items required for service of one type-l customer
N	the threshold of type-2 customers admission control
$(\boldsymbol{\beta}_l, S_l)$, $l = 1, 2$	an irreducible representation of the type-l customer service time distribution
M_l, $l = 1, 2$	the number transient states of the PH service process of type-l customer
$\mathbf{s}_0^{(l)}$, $l = 1, 2$	the transition intensities to the absorbing state of the PH service process of type-l customer
α	the intensity of impatience of type-2 customers from the buffer

The Markov chain $\xi_t,\ t \geq 0$, has the following state space:

$$\begin{aligned} &\Big(\{0, 0, k, \nu, \zeta\},\ k = \overline{0, K}\Big) \\ &\quad \cup \Big(\{i, 0, k, \nu, \zeta\},\ i > 0, k = \overline{0, N}\Big) \\ &\quad \cup \Big(\{i, r, k, \nu, \zeta, \eta\},\ i \geq 0,\ r = 1, 2,\ k = \overline{0, K}\Big), \\ &\nu = \overline{0, W}, \quad \zeta = \overline{0, V}, \quad \eta = \overline{1, \delta_{r=1} M_1 + \delta_{r=2} M_2}. \end{aligned}$$

For further use throughout this paper, we introduce the following notation:

- I is the identity matrix, and O is the zero matrix of the appropriate dimension;
- $\otimes$ and $\oplus$ indicate the symbols of Kronecker's product and sum of matrices, respectively;
- $\bar{W} = W + 1$, $\bar{V} = V + 1$;
- $\mathrm{diag}\{A_1, \ldots, A_l\}$ is a block-diagonal matrix with the diagonal blocks $A_1, \ldots, A_l$;
- E_N^+ is the square matrix of size $N + 1$ with all zero entries except the entries $(E_N^+)_{l,l+1},\ l = \overline{0, N-1}$, which are equal to 1;
- E_K^+ is the square matrix of size $K + 1$ with all zero entries except the entries $(E_K^+)_{l,l+1},\ l = \overline{0, K-1}$, and $(E_K^+)_{K,K}$ which are equal to 1;
- $E_{n,k}^-,\ n = N, K,\ k = k_1, k_2$, is the $(n + 1) \times (K+1)$ matrix with all zero entries except the entries $(E_{n,k}^-)_{l,l-k},\ l = \overline{k, n}$, which are equal to 1;
- $\tilde{E}$ is the $(N+1)\times(K+1)$ matrix with all zero entries except the entry $(\tilde{E})_{N,N-k_2+1}$ which is equal to 1;
- $\bar{E}$ is the square matrix of size $K + 1$ with all zero entries except the entries $(\bar{E})_{l,l-k_2},\ l = \overline{N+1, K}$, which are equal to 1;
- $\bar{I}_{n,k},\ n = N, K,\ k = k_1, k_2$, is the square matrix of size $n + 1$ with all zero entries except the entries $(\bar{I}_{n,k})_{l,l},\ l = \overline{0, \min\{k-1, n\}}$, which are equal to 1;
- $I_{k,j},\ k, j \geq 0$, is the $(k + 1) \times (j + 1)$ matrix with all zero entries except for the entries $(I_{k,j})_{l,l},\ l = \overline{0, \min\{k, j\}}$, which are equal to 1.

Let us enumerate the states of the Markov chain $\xi_t,\ t \geq 0$, in the direct lexicographic order of the components r, k, ν, ζ, η and refer to the set of the states of the chain having values (i, r) of the first two components of the Markov chain as a macro-state (i, r). Let Q be the generator of the Markov chain $\xi_t,\ t \geq 0$, consisting of the blocks $Q_{i,j}$, which, in turn, consist of the

matrices $Q_{i,j}^{(r,r')}$ of the transition rates of this chain from the macro-state (i,r) to the macro-state (j,r'), $r,r' = 0,1,2$. The diagonal entries of the matrices $Q_{i,i}$ are negative, and the modulus of the diagonal entry of the blocks $Q_{i,i}^{(r,r)}$ defines the total intensity of leaving the corresponding state of the Markov chain ξ_t, $t \geq 0$.

Lemma 1. *The infinitesimal generator $Q = (Q_{i,j})_{i,j\geq 0}$ of the Markov chain ξ_t, $t \geq 0$, has a block-tridiagonal structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \dots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & \dots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The non-zero blocks $Q_{i,j}$, $i,j \geq 0$, have the following form:

$$Q_{i,i} = \begin{pmatrix} Q_{i,i}^{(0,0)} & Q_{i,i}^{(0,1)} & Q_{i,i}^{(0,2)} \\ Q_{i,i}^{(1,0)} & Q_{i,i}^{(1,1)} & O \\ Q_{i,i}^{(2,0)} & Q_{i,i}^{(2,1)} & Q_{i,i}^{(2,2)} \end{pmatrix}, \quad i \geq 0, \quad (1)$$

$$\begin{aligned} Q_{i,i}^{(0,0)} &= I_{\delta_{i=0}K+\delta_{i>0}N+1} \otimes (D_0 \oplus H_0) \\ &\quad + E^{+}_{\delta_{i=0}K+\delta_{i>0}N} \otimes I_{\bar{W}} \otimes H_1 - i\alpha I \\ &\quad + \bar{I}_{\delta_{i=0}K+\delta_{i>0}N,k_1} \otimes D_1 \otimes I_{\bar{V}}, \quad i \geq 0, \end{aligned}$$

$$Q_{i,i}^{(0,1)} = E^{-}_{\delta_{i=0}K+\delta_{i>0}N,k_1} \otimes D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}_1, \quad i \geq 0,$$

$$Q_{i,i}^{(0,2)} = \begin{cases} \bar{E} \otimes D_2 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}_2, & i = 0, \\ O, & i > 0, \end{cases}$$

$$\begin{aligned} Q_{i,i}^{(r,r)} &= I_{K+1} \otimes (D_0 \oplus H_0 \oplus S_r) \\ &\quad + E_K^{+} \otimes I_{\bar{W}} \otimes H_1 \otimes I_{M_r} \\ &\quad - i\alpha I + (\delta_{r=2}\bar{I}_{K,k_1} + \delta_{r=1}I_{K+1}) \\ &\quad \otimes D_1 \otimes I_{\bar{V}M_r}, \quad r = 1,2, \end{aligned}$$

$$Q_{i,i}^{(2,1)} = (E^{-}_{K,k_1} \otimes D_1 \otimes I_{\bar{V}}) \otimes \mathbf{e}_{M_2}\boldsymbol{\beta}_1,$$

$$Q_{i,i}^{(r,0)} = \begin{cases} I_{(K+1)\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(r)}, & i = 0, \\ I_{K,N} \otimes I_{\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(r)}, & i > 0, \end{cases}$$

$$Q_{i,i+1} = \operatorname{diag}\{Q_{i,i+1}^{(r,r)}, r = \overline{0,2}\}, \quad i \geq 0, \quad (2)$$

$$Q_{i,i+1}^{(0,0)} = \begin{cases} I_{K,N} \otimes D_2 \otimes I_{\bar{V}}, & i = 0, \\ I_{N+1} \otimes D_2 \otimes I_{\bar{V}}, & i > 0, \end{cases}$$

$$Q_{i,i+1}^{(r,r)} = I_{K+1} \otimes D_2 \otimes I_{\bar{V}M_r}, \quad r = 1,2,$$

$$Q_{i,i-1} = \begin{pmatrix} Q_{i,i-1}^{(0,0)} & O & Q_{i,i-1}^{(0,2)} \\ O & Q_{i,i-1}^{(1,1)} & Q_{i,i-1}^{(1,2)} \\ O & O & Q_{i,i-1}^{(2,2)} \end{pmatrix}, \quad i \geq 1, \quad (3)$$

$$Q_{i,i-1}^{(r,r)} = \begin{cases} i\alpha I_{N,K} \otimes I_{\bar{W}\bar{V}}, & r = 0, i = 1, \\ i\alpha I_{(N+1)\bar{W}\bar{V}}, & r = 0, i > 1, \\ i\alpha I_{(K+1)\bar{W}\bar{V}M_1}, & r = 1, i > 0, \\ i\alpha I_{(K+1)\bar{W}\bar{V}M_2} & \\ \quad + \bar{E} \otimes I_{\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(2)}\boldsymbol{\beta}_2, & r = 2, i > 0, \end{cases}$$

$$Q_{i,i-1}^{(0,2)} = \tilde{E} \otimes I_{\bar{W}} \otimes H_1 \otimes \boldsymbol{\beta}_2,$$

$$Q_{i,i-1}^{(1,2)} = \bar{E} \otimes I_{\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(1)}\boldsymbol{\beta}_2.$$

Proof. To give insight into derivation of the generator of a multi-dimensional Markov chain for the model considered and give a possibility to an interested reader to analyze possible modifications of our model, we present a brief explanation of the form of the generator and its blocks.

The generator Q has a block-tridiagonal structure ($Q_{i,j} = O$ if $|i-j| > 1$) because the probability of two or more arrivals or departures (as well as the simultaneous arrival and departure) during the small interval is negligible. All non-zero matrices $Q_{i,j}$ consist of nine blocks $(Q_{i,j})^{(r,r')}$, $r,r' = 0,1,2$, containing the intensities of transition of the Markov chain from the macro-state (i,r) to the macro-state (j,r'), $r,r' = 0,1,2$.

In particular, this implies that the matrix $Q_{i,i}$, $i \geq 0$, can be presented in the form (1). The zero block in (1) appears since there is no transition from the macro-state $(i,1)$ to the macro-state $(i,2)$ because service of the priority customer cannot be terminated by a non-priority customer. The negative diagonal entries of the matrices $(Q_{i,i})^{(r,r)}$, $r = 0,1,2$, define, up to the sign, the intensities of the exit of the Markov chain ξ_t from the macro-state (i,r). Such an exit can happen due to a change in the state of the underlying processes of the MMAP or the MAP arrival processes, a customer service completion if $r = 1$ and $r = 2$ and the departure of a customer from the buffer due to impatience. The non-diagonal entries of the matrices $(Q_{i,i})^{(r,r)}$, $r = 0,1,2$, define the intensities of transition of the Markov chain ξ_t that do not lead to the change in the macro-state (i,r): the change in the state of the underlying process of the MAP arrivals of additional items, which does not cause the start of service of the waiting type-2 customer, the change in the state of the underlying process of the MMAP arrival of customers that does not lead to a customer generation (non-diagonal entries of the matrix D_0), the change in the state of the PH service process that does not lead to the service completion in case $r = 1$ and $r = 2$ (non-diagonal entries of the matrix S_r) and the arrival and rejection of a priority customer.

The matrices $(Q_{i,i})^{(0,r)}$, $r = 1,2$, define the intensities of admission of the type-r customer for service upon arrival. In this case, the corresponding number of additional items disappears and the service process is

started. The matrices $(Q_{i,i})^{(r,0)}$, $r = 1, 2$, define the intensities of the event when a type-r customer completes service and new service does not start. The matrix $(Q_{i,i})^{(2,1)}$ defines the intensities of the event when service of type-2 customer is terminated by the arrival of a priority customer.

The matrix $Q_{i,i+1}$, $i \geq 0$, defines the intensities of the events that lead to an increase in the number of type-2 customers in the buffer and has form (2). The diagonal form of (2) is explained by the fact that the events that lead to an increase in the number of type-2 customers in the buffer do not change the state of the server. Accordingly, the non-diagonal blocks of the matrix $Q_{i,i+1}$, $i \geq 0$, have all zero entries. The block $Q_{i,i+1}^{(r,r)}$, $i \geq 0, r = 0, 1, 2$, defines the intensities of type-2 customer arrival when this customer cannot immediately start service due to the lack of additional items or server business.

The matrix $Q_{i,i-1}$, $i \geq 1$, defines the intensities of the events that lead to a decrease in the number of type-2 customers in the buffer and has form (3). The blocks $Q_{i,i-1}^{(0,1)}$, $Q_{i,i-1}^{(1,0)}$, $Q_{i,i-1}^{(2,0)}$, and $Q_{i,i-1}^{(2,1)}$, are equal to zero matrices because the corresponding changes in the state of the server are impossible when the number of type-2 customers in the buffer decreases. The blocks $(Q_{i,i-1})^{(r,r)}$, $r = 0, 1, 2$, define the intensities of type-2 customers leaving the buffer due to impatience. Also the block $(Q_{i,i-1})^{(2,2)}$ defines the intensity of the event that service of a type-2 customer is finished and a type-2 customer from the buffer is chosen for service. The block $(Q_{i,i-1})^{(0,2)}$ defines the intensity of the event that an item arrives to the system when there are N items in the buffer and type-2 customer from the buffer is admitted for service. The block $(Q_{i,i-1})^{(1,2)}$ defines the intensity of the event that service of a type-1 customer is finished and a type-2 customer from the buffer starts service. ■

Remark 1. As was already mentioned above, the understanding of the generator gives an opportunity to analyze possible modifications of our model. For example, for some applications the loss of a type-2 customer in the case of the arrival of a type-1 customer may be unrealistic. Modification of the generator to the case when the customer whose service is interrupted is not lost but returns to the queue can be easily performed using our results. To this end, it is necessary to make the following two modifications: (i) the block $Q_{i,i}^{(2,1)}$ of the matrix $Q_{i,i}$ should be zero block; (ii) the block $Q_{i,i+1}^{(2,1)}$ of the matrix $Q_{i,i+1}$ should be given by the formula $Q_{i,i+1}^{(2,1)} = (E_{K,k_1}^{-} \otimes D_1 \otimes I_{\bar{V}}) \otimes \mathbf{e}_{M_2}\boldsymbol{\beta}_1$.

Remark 2. The Markov chain ξ_t, $t \geq 0$, belongs to the class of continuous-time asymptotically quasi-Toeplitz Markov chains (AQTMC) (see Klimenok and Dudin, 2006).

Let us analyze the properties of this Markov chain. This analysis should include derivation of conditions, which should be imposed on the system parameters to guarantee existence of a stationary distribution of the states of the chain (the ergodicity condition), and a procedure for computation of the stationary probabilities of the states.

It follows from the work of Klimenok and Dudin (2006) that a sufficient condition for the existence of a stationary distribution of $AQTMC$ ξ_t, $t \geq 0$, can be expressed in terms of the matrices Y_0, Y_1 and Y_2 defined as follows:

$$Y_0 = \lim_{i\to\infty} R_i^{-1} Q_{i,i-1},$$

$$Y_1 = \lim_{i\to\infty} R_i^{-1} Q_{i,i} + I,$$

$$Y_2 = \lim_{i\to\infty} R_i^{-1} Q_{i,i+1},$$

where the matrix R_i is the diagonal matrix with the diagonal entries which are defined as the moduli of the corresponding diagonal entries of the matrix $Q_{i,i}$, $i \geq 0$.

It is easy to verify that in the case considered the matrices Y_0, Y_1 and Y_2 have the following form:

$$Y_0 = I, \quad Y_1 = O, \quad Y_2 = O,$$

and, as it follows from the work of Klimenok and Dudin (2006), a sufficient condition for the ergodicity of Markov chain ξ_t, $t \geq 0$, is the fulfillment of the inequality

$$\mathbf{y}Y_0\mathbf{e} > \mathbf{y}Y_2\mathbf{e},$$

where the vector $\mathbf{y}$ is the unique solution to the system

$$\mathbf{y}(Y_0 + Y_1 + Y_2) = \mathbf{y}, \quad \mathbf{y}\mathbf{e} = 1.$$

It is easy to see that here the ergodicity condition is given by the inequality $1 > 0$ which is true for all possible values of the system parameters. The intuitive explanation of this fact is clear. Because the customers in the buffer are impatient and leave the system after an exponentially distributed amount of time, the number of customers in the buffer never approaches infinity.

Accordingly, the stationary probabilities of the system states $\pi(i, r, k, \nu, \zeta, \eta)$, $i \geq 0$, $r = \overline{0,2}$, $k = \overline{0,K}$, $\nu = \overline{0,W}$, $\zeta = \overline{0,V}$, $\eta = \overline{1,M_r}$, always exist. Let us form the row vectors $\boldsymbol{\pi}(i, r, k)$ of these probabilities enumerated in the lexicographic order of the components ν, ζ, η. Then let us form the row vectors

$$\boldsymbol{\pi}(0,0) = (\boldsymbol{\pi}(0,0,0), \boldsymbol{\pi}(0,0,1), \ldots, \boldsymbol{\pi}(0,0,K)),$$

$$\boldsymbol{\pi}(i,0) = (\boldsymbol{\pi}(i,0,0), \boldsymbol{\pi}(i,0,1), \ldots, \boldsymbol{\pi}(i,0,N)), \quad i \geq 1,$$

$$\boldsymbol{\pi}(i,r) = (\boldsymbol{\pi}(i,r,0), \boldsymbol{\pi}(i,r,1), \dots, \boldsymbol{\pi}(i,r,K)), \quad r = 1,2,$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \boldsymbol{\pi}(i,2)), \quad i \geq 0.$$

It is well known that the probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, satisfy the following system of linear algebraic equations:

$$\begin{aligned}(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)Q &= \mathbf{0},\\ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)\mathbf{e} &= 1,\end{aligned} \tag{4}$$

where Q is the infinitesimal generator of the Markov chain ξ_t. The system of equations (4) is infinite and the problem of its solution is quite difficult. However, it can be successfully solved using the numerically stable algorithm that takes into account that the matrix Q has a block-tridiagonal structure and the limits $Y_k, k = 0,1,2$, exist, which is presented by Dudina *et al.* (2013).

For convenience, we present this algorithm taking into account the features of the generator Q (see Algorithm 1).

4. Performance measures

As soon as the vectors $\boldsymbol{\pi}_i$, $i \geq 0$, have been calculated, we are able to find various performance measures of the system.

The probability P_1^{serv} that at an arbitrary epoch the server is busy by a type-1 customer is computed by

$$P_1^{\text{serv}} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,1)\mathbf{e}.$$

Remark 3. The formulas for computing the main performance of the system contain infinite sums. The question about calculation of such sums arises. Note that the Markov chain that describes the system behavior is ergodic; therefore, the stationary probability vectors $\boldsymbol{\pi}_i$ converge in norm to a zero vector as i approaches infinity. Thus, the computation of some infinite sum may be terminated if the norm of the summand becomes less than a preassigned value ε (e.g., $\varepsilon = 10^{-10}$).

The probability P_2^{serv} that at an arbitrary epoch the server is busy by a type-2 customer is computed by

$$P_2^{\text{serv}} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,2)\mathbf{e}.$$

The average number $N_{\text{customers}}^{\text{buffer}}$ of customers in the buffer is computed by

$$N_{\text{customers}}^{\text{buffer}} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}.$$

Algorithm 1.

Step 1. Calculate the stochastic matrices G_i using the recursion

$$G_i = -(Q_{i+1,i+1} + Q_{i+1,i+2}G_{i+1})^{-1}Q_{i+1,i}, \quad i \geq 0.$$

Note that this recursion is backwards and for computing the matrix G_i it is necessary to obtain all matrices G_l, $l > i$, $i \geq 0$. It can be proven that in the case considered the sequence of the matrices G_i converges to the identity matrix, when i approaches infinity. Thus, for any predefined small positive number ϵ_G there exists a value i_0 such that the norm of the matrix $G_i - I$ is less than ε_G for all l, $l \geq i_0$. Thus, we can set in the backward recursion $G_l = I$ for l, $l \geq i_0$.

Step 2. Calculate the matrices F_i using the recursion

$$F_i = -F_{i-1}Q_{i-1,i}(Q_{i,i} + Q_{i,i+1}G_i)^{-1}, \quad i \geq 1,$$

with the initial condition $F_0 = I$.

Note that the norm of the matrix F_i tends to zero when i approaches infinity. Thus, the calculation of the matrices F_i can be terminated if the norm of the matrix F_i becomes less than some preassigned positive value.

Step 3. Calculate the vector $\boldsymbol{\pi}_0$ as the unique solution to the system of the linear algebraic equations

$$\boldsymbol{\pi}_0(Q_{0,0} + Q_{0,1}G_0) = \mathbf{0}, \quad \boldsymbol{\pi}_0 \sum_{i=0}^{\infty} F_i\mathbf{e} = 1.$$

Step 4. Calculate the stationary probabilities vectors $\boldsymbol{\pi}_i$, $i \geq 1$, as

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 F_i, \quad i \geq 1.$$

The average number $N_{\text{item}}^{\text{stock}}$ of additional items in the stock is computed by

$$N_{\text{item}}^{\text{stock}} = \sum_{k=1}^{K} k\boldsymbol{\pi}(0,0,k)\mathbf{e} + \sum_{i=1}^{\infty}\sum_{k=1}^{N} k\boldsymbol{\pi}(i,0,k)\mathbf{e} + \sum_{i=0}^{\infty}\sum_{r=1}^{2}\sum_{k=1}^{K} k\boldsymbol{\pi}(i,r,k)\mathbf{e}.$$

The average intensity λ_l^{out} of flow of type-l customers who receive service is computed by

$$\lambda_l^{\text{out}} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,l)(\mathbf{e}_{(K+1)\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(l)}), \quad l = 1,2.$$

The probability P_l^{loss} that an arbitrary type-l customer will be lost is computed by

$$P_l^{\text{loss}} = 1 - \frac{\lambda_l^{\text{out}}}{\lambda_l}, \quad l = 1,2.$$

The probability $P_1^{\text{server-busy}}$ of an arbitrary arriving type-1 customer loss because the server is busy with a type-1 customer is computed as

$$P_1^{\text{server-busy}} = \frac{1}{\lambda_1}\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,1)(I_{K+1}\otimes D_1\otimes I_{\bar{V}M_1})\mathbf{e}.$$

The probability $P_1^{\text{item-lack}}$ of an arbitrary arriving type-1 customer loss due to the lack of additional items is computed as

$$\begin{aligned}P_1^{\text{item-lack}} &= \frac{1}{\lambda_1}\Bigg[\sum_{k=0}^{k_1-1}\boldsymbol{\pi}(0,0,k)(D_1\otimes I_{\bar{V}})\mathbf{e}\\ &\quad+\sum_{i=1}^{\infty}\sum_{k=0}^{\min\{k_1-1,N\}}\boldsymbol{\pi}(i,0,k)(D_1\otimes I_{\bar{V}})\mathbf{e}\\ &\quad+\sum_{i=0}^{\infty}\sum_{k=0}^{k_1-1}\boldsymbol{\pi}(i,2,k)(D_1\otimes I_{\bar{V}M_2})\mathbf{e}\Bigg].\end{aligned}$$

The probability P^{term} that service of an arbitrary type-2 customer is terminated by an arriving type-1 customer is computed as

$$P^{\text{term}} = \frac{1}{\lambda_2}\sum_{i=0}^{\infty}\sum_{k=k_1}^{K}\boldsymbol{\pi}(i,2,k)(D_1\otimes I_{\bar{V}M_2})\mathbf{e}.$$

The probability P^{imp} that an arbitrary type-2 customer leaves the system due to impatience is computed as

$$P^{\text{imp}} = P_2^{\text{loss}} - P^{\text{term}}.$$

The probability $P_{\text{item}}^{\text{loss}}$ that an arbitrary additional item will be lost is computed by

$$\begin{aligned}P_{\text{item}}^{\text{loss}} = \frac{1}{\lambda_e}\Bigg[&\sum_{i=1}^{\infty}\sum_{r=1}^{2}\boldsymbol{\pi}(i,r,K)(I_{\bar{W}}\otimes H_1\otimes I_{M_r})\mathbf{e}\\ &+\boldsymbol{\pi}(0,0,K)(I_{\bar{W}}\otimes H_1)\mathbf{e}\Bigg].\end{aligned}$$

The probability P_1^{imm} that an arbitrary arriving type-1 customer occupies the server is computed as

$$\begin{aligned}P_1^{\text{imm}} = \frac{1}{\lambda_1}\sum_{i=0}^{\infty}\Bigg[&\sum_{k=k_1}^{\delta_{i=0}K+\delta_{i>0}N}\boldsymbol{\pi}(i,0,k)(D_1\otimes I_{\bar{V}})\mathbf{e}\\ &+\sum_{k=k_1}^{K}\boldsymbol{\pi}(i,2,k)(D_1\otimes I_{\bar{V}M_2})\mathbf{e}\Bigg].\end{aligned}$$

The probability P_2^{imm} that at an arbitrary type-2 customer occupies the server upon arrival is computed as

$$P_2^{\text{imm}} = \frac{1}{\lambda_2}\sum_{k=N+1}^{K}\boldsymbol{\pi}(0,0,k)(D_2\otimes I_{\bar{V}})\mathbf{e}.$$

5. Distribution of the waiting time in the system of an arbitrary non-priority customer

We will derive the distribution of an arbitrary type-2 customer's waiting time in terms of the Laplace–Stieltjes transform (LST). Let $v(s)$ be the *LST* of the distribution of an arbitrary type-2 customer's waiting time. To derive the expression for this *LST*, we use the method of collective marks (method of additional event, method of catastrophes) (for references, see, e.g., Kesten and Runnenburg, 1956; Dantzig, 1955). To this end, we interpret the variable s as the intensity of some virtual stationary Poisson flow of catastrophes. In consequence, $v(s)$ has the meaning of the probability that no catastrophe happens during the waiting time of an arbitrary type-2 customer. Let us tag an arbitrary type-2 customer and keep track of its staying in the system. Accordingly, $v(s)$ has the meaning of the probability that the catastrophe does not arrive during the waiting time of the tagged type-2 customer.

Let $v(s,l,r,k,\nu,\zeta,\eta)$ be the probability that a catastrophe will not arrive during the rest of the tagged type-2 customer's waiting time in the system conditioned on the fact that at the given moment the position of the tagged customer in the buffer is $l,\ l\geq 1$, the state of the server is $r,\ r=\overline{0,2}$, the number of additional items in the stock is equal to $k,\ k=\overline{0,K}$, the states of the processes ν_t, ζ_t and η_t are ν, ζ and η respectively, $t\geq 0$.

Let us enumerate the probabilities $v(s,l,r,k,\nu,\zeta,\eta)$ in the lexicographic order of the components k,ν,ζ,η and form the column vectors $\mathbf{v}(s,l,r)$ from these probabilities.

To compute the unknown vectors $\mathbf{v}(s,l,r),\ r=0,1,2,\ l\geq 1$, let us introduce the column vectors

$$\mathbf{v}(s,l)=(\mathbf{v}^T(s,l,0),\mathbf{v}^T(s,l,1),\mathbf{v}^T(s,l,2))^T,\quad l\geq 1.$$

Theorem 1. *The vectors $\mathbf{v}(s,l),\ l\geq 1$, can be computed from the following recursion:*

$$\mathbf{v}(s,1)=((s+\alpha)I-V_1)^{-1}(\mathbf{a}+\alpha\mathbf{e}), \tag{5}$$

$$\begin{aligned}\mathbf{v}(s,l)=((s+l\alpha)I-V_1)^{-1}&((V_2+(l-1)\alpha I)\\ &\times\mathbf{v}(s,l-1)+\alpha\mathbf{e}),\quad l>1,\end{aligned} \tag{6}$$

where the matrices V_1 and V_2 and the vector $\mathbf{a}$ are defined as follows:

$$V_1=\begin{pmatrix}V_1^{(0,0)} & V_1^{(0,1)} & O\\ V_1^{(1,0)} & V_1^{(1,1)} & O\\ V_1^{(2,0)} & V_1^{(2,1)} & V_1^{(2,2)}\end{pmatrix},$$

$$\begin{aligned}V_1^{(0,0)} &= I_{N+1}\otimes((D_0+D_2)\oplus H_0)\\ &\quad+E_N^{+}\otimes I_{\bar{W}}\otimes H_1+\bar{I}_{N,k_1}\otimes D_1\otimes I_{\bar{V}},\end{aligned}$$

$$V_1^{(0,1)} = E^-_{N,k_1} \otimes D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}_1,$$

$$\begin{aligned} V_1^{(r,r)} = {} & I_{K+1} \otimes ((D_0 + D_2) \oplus H_0 \oplus S_r) \\ & + E_K^+ \otimes I_{\bar{W}} \otimes H_1 \otimes I_{M_r} \\ & + (\delta_{r=2}\bar{I}_{K,k_1} + \delta_{r=1} I_{K+1}) \\ & \otimes D_1 \otimes I_{\bar{V}M_r}, \quad r = 1, 2, \end{aligned}$$

$$V_1^{(2,1)} = (E^-_{K,k_1} \otimes D_1 \otimes I_{\bar{V}}) \otimes \mathbf{e}_{M_2} \boldsymbol{\beta}_1,$$

$$V_1^{(r,0)} = I_{K,N} \otimes I_{\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(r)}, \quad r = 1, 2,$$

$$V_2 = \begin{pmatrix} O & O & V_2^{(0,2)} \\ O & O & V_2^{(1,2)} \\ O & O & V_2^{(2,2)} \end{pmatrix},$$

$$V_2^{(0,2)} = \tilde{E} \otimes I_{\bar{W}} \otimes H_1 \otimes \boldsymbol{\beta}_2,$$

$$V_2^{(r,2)} = \bar{E} \otimes I_{\bar{W}\bar{V}} \otimes (\mathbf{s}_0^{(r)} \boldsymbol{\beta}_2), \quad r = 1, 2,$$

$$\begin{aligned} \mathbf{a} = {} & (((\tilde{E} \otimes I_{\bar{W}} \otimes H_1)\mathbf{e})^T, (\bar{E} \otimes I_{\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(1)})\mathbf{e})^T, \\ & ((\bar{E} \otimes I_{\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(2)})\mathbf{e})^T)^T. \end{aligned}$$

Proof. Through on analysis of all possible transitions of the components $l, r, k, \nu, \zeta, \eta$ during the infinitesimally small time interval and using the formula of total probability, it can be shown that the vectors $\mathbf{v}(s, l, r)$ can be found from the following system of linear algebraic equations:

$$\begin{aligned} & \mathbf{v}(s, l, 0) \\ & = \Big[(s + l\alpha)I - I_{N+1} \otimes (D_0 \oplus H_0)\Big]^{-1} \\ & \quad \times \Big(\delta_{l>1} \tilde{E} \otimes I_{\bar{W}} \otimes H_1 \otimes \boldsymbol{\beta}_2 \mathbf{v}(s, l-1, 2) \\ & \quad + \delta_{l=1} \tilde{E} \otimes I_{\bar{W}} \otimes H_1 \mathbf{e} + (I_{N+1} \otimes D_2 \otimes I_{\bar{V}} \\ & \quad + \bar{I}_{N,k_1} \otimes D_1 \otimes I_{\bar{V}} + E_N^+ \otimes I_{\bar{W}} \otimes H_1)\mathbf{v}(s, l, 0) \\ & \quad + E^-_{N,k1} \otimes D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}_1 \mathbf{v}(s, l, 1) \\ & \quad + (l-1)\alpha \mathbf{v}(s, l-1, 0) + \alpha \mathbf{e}\Big), \end{aligned} \tag{7}$$

$$\begin{aligned} & \mathbf{v}(s, l, r) \\ & = \Big[(s + l\alpha)I - I_{N+1} \otimes (D_0 \oplus H_0 \oplus S_r)\Big]^{-1} \\ & \quad \times \Big(\big(((\delta_{r=1} I_{K+1} + \delta_{r=2}\bar{I}_{K,k_1}) \otimes D_1 \\ & \quad + I_{K+1} \otimes D_2) \otimes I_{\bar{V}M_r} \\ & \quad + E_K^+ \otimes I_{\bar{W}} \otimes H_1 \otimes I_{M_r})\mathbf{v}(s, l, r) \\ & \quad + \delta_{l>1} \bar{E} \otimes I_{\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(r)} \boldsymbol{\beta}_2 \mathbf{v}(s, l-1, 2) \\ & \quad + \delta_{l=1}(\bar{E} \otimes I_{\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(r)})\mathbf{e} \\ & \quad + \delta_{r=2} E^-_{K,k_1} \otimes D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}_1 \mathbf{v}(s, l, 1) \\ & \quad + (l-1)\alpha \mathbf{v}(s, l-1, r) + \alpha \mathbf{e} \\ & \quad + I_{K,N} \otimes I_{\bar{W}\bar{V}} \otimes \mathbf{s}_0^{(r)} \mathbf{v}(s, l, 0)\Big), \quad r = 1, 2,\ l \geq 1. \end{aligned} \tag{8}$$

Let us briefly explain (7). The term in the square bracket of the right-hand side of (7) defines the total intensity of the events which can happen after an arbitrary time moment: a catastrophe arrival, the transition of the MMAP arrival process of customers, the transition of the MAP arrival process of additional items, a type-2 customer leaving from the buffer. Note that in the case considered the server is free ($r = 0$). Therefore the transitions of the service process are impossible. The first term in the round brackets in (7) corresponds to the case when the tagged customer is not the first in the queue, the number of additional items in the buffer is $N - 1$ and the new additional item arrives to the system. In this situation, the first type-2 customer from the queue starts service and the position of the tagged customer in the buffer decreases by one. Thus, the conditional probability that a catastrophe will not arrive during the rest of the tagged type-2 customer's waiting time is defined as the corresponding entry of the vector $\mathbf{v}(s, l-1, 2)$. The second term corresponds to the situation when the tagged customer is the first in the queue, the number of additional items in the buffer is $N-1$ and a new additional item arrives to the system. In this situation, the tagged customer starts service. In this case, the waiting time of the tagged customer is finished and the conditional probability that a catastrophe will not arrive is defined as the corresponding entry of the vector $\mathbf{e}$.

The third term in the square bracket corresponds to the situation when a new type-2 customer arrives to the system. The fourth term corresponds to the situation when the number of additional items in the buffer is less than k_1 and a new type-1 customer arrives, it is not admitted to the system and leaves it forever. The fifth term corresponds to the case when the number of additional items in the buffer is less than $N - 1$, a new additional item arrives and joins the buffer. In the third, fourth and fifth situations the position of the tagged customer and the state of the server are not changed. Thus, the conditional probability that a catastrophe will not arrive during the rest of the tagged customer's waiting time is defined as the corresponding entry of the vector $\mathbf{v}(s, l, 0)$. The sixth term corresponds to the situation when the number of additional items in the buffer is greater than or equal to k_1 and a new type-1 customer arrives. This customer starts service and the conditional probability that a catastrophe will not arrive during the rest of the tagged customer's

waiting time is defined as the corresponding entry of the vector $\mathbf{v}(s,l,1)$. The seventh term corresponds to the case when some non-tagged type-2 customer leaves the buffer due to impatience. In this case, the position of the tagged customer in the buffer decreases by one and the probability that a catastrophe will not arrive during the rest of the tagged type-2 customer's waiting time is defined as the corresponding entry of the vector $\mathbf{v}(s,l-1,0)$. Eventually, the eighth term corresponds to the situation when the tagged customer leaves the buffer due to impatience. The waiting time of the tagged customer is finished and the probability that a catastrophe will not arrive is defined as the corresponding entry of the vector $\mathbf{e}$.

The formula (8) is explained analogously. The only difference is that the server is not free. Therefore, the transitions of the service process should be taken into account.

To find the solution to the system (7) and (8), rewrite it in matrix form as

$$\begin{aligned}(-(s+l\alpha)I+V_1)\mathbf{v}(s,l)&\\ +\delta_{l>1}(V_2+(l-1)\alpha I)\mathbf{v}(s,l-1)&\\ +\delta_{l=1}\mathbf{a}+\alpha\mathbf{e}=\mathbf{0}^T,\quad l\geq 1.&\quad (9)\end{aligned}$$

The recursion (6) with the initial condition (5) evidently follows from (9). ■

Theorem 2. *The LST $v(s)$ of the distribution of an arbitrary type-2 customer's waiting time in the system is*

$$\begin{aligned}v(s)&\\ =P_2^{\text{imm}}+\lambda_2^{-1}\Bigg[&\sum_{k=0}^{N}\boldsymbol{\pi}(0,0,k)(D_2\otimes I_{\bar{V}})\mathbf{v}(s,1,0)\\ &+\sum_{i=1}^{\infty}\boldsymbol{\pi}(i,0)(I_{N+1}\otimes D_2\otimes I_{\bar{V}})\mathbf{v}(s,i+1,0)\\ &+\sum_{i=0}^{\infty}\sum_{r=1}^{2}\boldsymbol{\pi}(i,r)(I_{K+1}\otimes D_2\otimes I_{\bar{V}M_r})\mathbf{v}(s,i+1,r)\Bigg].\end{aligned}$$

Proof. Let us consider all possible situations at the arrival epoch of the tagged customer:

- The tagged customer starts service upon arrival. The probability of this event is P_2^{imm}. In this case, the probability that no catastrophe arrives during the waiting time is equal to one.
- The tagged customer arrives to the empty system, but the number of additional items is not sufficient to start service. The probability of this event is $\lambda_2^{-1}\sum_{k=0}^{N}\boldsymbol{\pi}(0,0,k)(D_2\otimes I_{\bar{V}})\mathbf{e}$. In this case, the conditional probability that no catastrophe arrives during the waiting time is equal to the corresponding entry of the vector $\mathbf{v}(s,1,0)$.
- The tagged customer arrives to the system when there are other type-2 customers in the system, but the server is free because the number of additional items is less than or equal to N. The probability of this event is $\lambda_2^{-1}\sum_{i=1}^{\infty}\boldsymbol{\pi}(i,0)(I_{N+1}\otimes D_2\otimes I_{\bar{V}})\mathbf{e}$. In this case, the conditional probability that no catastrophe arrives during the waiting time is equal to the corresponding entry of the vector $\mathbf{v}(s,i+1,0)$.
- The tagged customer arrives to the system when the server serves type-r customer. Its probability is $\lambda_2^{-1}\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,r)(I_{K+1}\otimes D_2\otimes I_{\bar{V}M_r})\mathbf{e}$. In this case, the conditional probability that no catastrophe arrives during the waiting time is equal to the corresponding entry of the vector $\mathbf{v}(s,i+1,r)$.

The assertion of the theorem evidently follows from the above considerations and the law of total probability. ■

Corollary 1. *The average waiting time V^{wait} of an arbitrary type-2 customer is*

$$\begin{aligned}V^{\text{wait}}&\\ =-\lambda_2^{-1}\Bigg[&\sum_{k=0}^{N}\boldsymbol{\pi}(0,0,k)(D_2\otimes I_{\bar{V}})\mathbf{v}'(s,1,0,k)|_{s=0}\\ &+\sum_{i=1}^{\infty}\boldsymbol{\pi}(i,0)(I_{N+1}\otimes D_2\otimes I_{\bar{V}})\mathbf{v}'(s,i+1,0)|_{s=0}\\ &+\sum_{i=0}^{\infty}\sum_{r=1}^{2}\boldsymbol{\pi}(i,r)(I_{K+1}\otimes D_2\otimes I_{\bar{V}M_r})\\ &\times\mathbf{v}'(s,i+1,r)|_{s=0}\Bigg].\end{aligned}$$

Here the column vectors $\mathbf{v}'(s,l,r)|_{s=0}$ and $\mathbf{v}'(s,l,r,k)|_{s=0}$ are calculated as the blocks of the vector $\mathbf{v}'(s,l)|_{s=0}$ which can be calculated recursively as follows:

$$\mathbf{v}'(s,1)|_{s=0}=-(\alpha I-V_1)^{-2}(\mathbf{a}+\alpha\mathbf{e}),$$

$$\begin{aligned}&\mathbf{v}'(s,l)|_{s=0}\\ &\quad=(l\alpha I-V_1)^{-1}(-\mathbf{e}+(V_2+(l-1)\alpha I)\\ &\qquad\times\mathbf{v}'(s,l-1)|_{s=0}),\quad l>1.\end{aligned}$$

Proof. The formula for calculation of the average waiting time of an arbitrary tagged customer is based on the definition $V^{\text{wait}}=-v'(s)|_{s=0}$. ■

On the analogy of Theorems 1 and 2, we can obtain the following assertion.

Theorem 3. *The LST $w(s)$ of the distribution of the waiting time of an arbitrary type-2 customer in the system*

that does not leave the system due to impatience is

$$
\begin{aligned}
&w(s)\\
&= ((1-P^{\text{imp}})\lambda_2)^{-1}\\
&\quad\times\Bigg[\sum_{k=0}^{N}\boldsymbol{\pi}(0,0,k)(D_2\otimes I_{\bar V})\mathbf{w}(s,1,0,k)\\
&\quad+\sum_{i=1}^{\infty}\boldsymbol{\pi}(i,0)(I_{N+1}\otimes D_2\otimes I_{\bar V})\mathbf{w}(s,i+1,0)\\
&\quad+\sum_{i=0}^{\infty}\sum_{r=1}^{2}\boldsymbol{\pi}(i,r)(I_{K+1}\otimes D_2\otimes I_{\bar V M_r})\mathbf{w}(s,i+1,r)\Bigg],
\end{aligned}
$$

where the column vectors $\mathbf{w}(s,l,r,k)$ *and* $\mathbf{w}(s,l,r)$ *are calculated as the blocks of the vector* $\mathbf{w}(s,l)$ *which can be determined from the following recursion:*

$$\mathbf{w}(s,1)=((s+\alpha)I-V_1)^{-1}\mathbf{a},$$

$$
\begin{aligned}
\mathbf{w}(s,l)&=((s+l\alpha)I-V_1)^{-1}\\
&\quad\times(V_2+(l-1)\alpha I)\mathbf{w}(s,l-1),\quad l>1.
\end{aligned}
$$

Corollary 2. *The average waiting time* $V^{\text{wait-patient}}$ *of an arbitrary type-2 customer that does not leave the system due to impatience is*

$$
\begin{aligned}
&V^{\text{wait-patient}}\\
&= -((1-P^{\text{imp}})\lambda_2)^{-1}\\
&\quad\times\Bigg[\sum_{k=0}^{N}\boldsymbol{\pi}(0,0,k)(D_2\otimes I_{\bar V})\mathbf{w}'(s,1,0,k)|_{s=0}\\
&\quad+\sum_{i=1}^{\infty}\boldsymbol{\pi}(i,0)(I_{N+1}\otimes D_2\otimes I_{\bar V})\mathbf{w}'(s,i+1,0)|_{s=0}\\
&\quad+\sum_{i=0}^{\infty}\sum_{r=1}^{2}\boldsymbol{\pi}(i,r)(I_{K+1}\otimes D_2\otimes I_{\bar V M_r})\\
&\quad\times\mathbf{w}'(s,i+1,r)|_{s=0}\Bigg].
\end{aligned}
$$

6. Numerical examples

Our numerical experiment has two goals. The first one is to show an effect of variation in the threshold N that defines the discipline of type-2 customers accepting to service. The second goal is to demonstrate the necessity of taking account of correlation in the arrival processes of customers and additional items.

We assume that the buffer of additional items (stock) has a capacity of $K=20$, the intensity of impatience of type-2 customers $\alpha=0.015$, the number of additional items required for service of one type-1 customer $k_1=5$ and the number of additional items required for service of one type-2 customer $k_2=2$. The motivation for the choice $k_1>k_2$ stems from the interpretation that a type-1 customer represents an emergency information unit, while a type-2 customer represents a routine information unit. Because the probability of the successful transmission of emergency information should be higher, this transmission should use a stronger signal, which requires more units of energy.

We assume that the PH service process of type-1 customers is characterized by the vector $\boldsymbol{\beta}_1=(1,0)$ and the matrix

$$S_1=\begin{pmatrix}-5 & 5\\ 0 & -5\end{pmatrix}.$$

The mean service time $b_1^{(1)}$ is equal to 0.4, the coefficient of variation is equal to 0.5. The PH service process of type-2 customers is characterized by the vector $\boldsymbol{\beta}_2=(1,0)$ and the matrix

$$S_2=\begin{pmatrix}-3 & 3\\ 0 & -3\end{pmatrix}.$$

The mean service time $b_1^{(2)}$ is equal to $2/3$, coefficient of variation is equal to 0.5. We fixed $b_1^{(2)}>b_1^{(1)}$ assuming that emergency information is shorter than routine information about the object state.

Aiming to demonstrate the effect of correlation, we introduce two MMAP arrival flows of customers and two MAP arrival flows of additional items. As is described above, the MMAP arrival flows of customers are defined by the matrices D_0, D_1 and D_2. We fix two MMAPs coded as MMAP^0 and MMAP$^{0.4}$. Both these MMAPs have the same average total arrival rate $\lambda=1$, the average intensity of priority customers $\lambda_1=0.1$, and the average intensity of non-priority customers $\lambda_2=0.9$, but different coefficients of correlation of successive inter-arrival times and variation. The MMAP^0 is defined by the matrices $D_0=-1$, $D_1=0.1$ and $D_2=0.9$. It has the coefficient of correlation $c_{cor}=0$ and the coefficient of variation $c_{var}=1$. The MMAP$^{0.4}$ is defined by the matrices

$$D_0=\begin{pmatrix}-3.3977 & 0\\ 0.001 & -0.1102\end{pmatrix},$$

$$D_1=\begin{pmatrix}0.3362 & 0.0035\\ 0.0012 & 0.0097\end{pmatrix},$$

$$D_2=\begin{pmatrix}3.026 & 0.032\\ 0.0109 & 0.0874\end{pmatrix}.$$

It has the coefficient of correlation $c_{cor}=0.4$, and the coefficient of variation $c_{var}=12.39$.

We also fix two MAP arrival flows of additional items defined by the matrices H_0 and H_1. Both these MAPs have the same average total arrival rate $\lambda_e=2.5$, but different coefficients of correlation. The arrival process of the first items coded as MAP^0 is defined by the matrices $H_0=-2.5$ and $H_1=2.5$. It has the coefficient of correlation $c_{cor}=0$ and the coefficient of variation

$c_{var} = 1$. The second process of items arrivals coded as MAP$^{0.4}$ is defined by the matrices

$$H_0 = \begin{pmatrix} -8.4942 & 0 \\ 0.0025 & -0.2755 \end{pmatrix},$$

$$H_1 = \begin{pmatrix} 8.4057 & 0.0885 \\ 0.0303 & 0.2427 \end{pmatrix}.$$

It has the coefficient of correlation $c_{cor} = 0.4$, and the coefficient of variation $c_{var} = 12.39$.

Let us vary the threshold N in the interval $[k_2 - 1, K - 1]$. Figure 8 illustrates the dependence of the loss probability P_1^{loss} of an arbitrary type-1 customer and the loss probability $P_{\text{item}}^{\text{loss}}$ of an arbitrary additional item on the threshold N.

It can be observed in this figure that the loss probability $P_{\text{item}}^{\text{loss}}$ is almost constant and only slightly increases when N grows. The increase is easily explained by the fact that when N is large, the non-priority customers have low chances to reach the server. Many non-priority customers are lost due to impatience. This implies smaller consumption of additional items and, as a consequence, the overflow of the stock and the loss of arriving items. It is worth noticing that $P_{\text{item}}^{\text{loss}}$ essentially depends on correlation in the arrival processes of customers and additional items. The absence of correlation implies a more uniform arrival of customers and additional items. High correlation causes alternation of periods when the customers and (or) items arrive frequently with periods when the customers and (or) items arrive rarely. This has negative impact on the value of $P_{\text{item}}^{\text{loss}}$.

The behavior of P_1^{loss} is quite complicated. Some dependencies are not monotone. The probability P_1^{loss} is pretty high for $N \leq 5$. When $N > 5$, this loss probability essentially decreases. This phenomenon is easily explained as follows. Because a type-1 customer may be accepted for service only with the presence of at least $k_1 = 5$ additional items, while any type-2 customer requires for service $k_2 = 2$ additional items, in the situation when $N \leq 5$ type-2 customers have good conditions for access to service and permanently expend the additional items. Accordingly, it is difficult to have $k_1 = 5$ additional items in the stock and type-1 customers rarely reach the server. As in case of $P_{\text{item}}^{\text{loss}}$, we observe again strong dependence of the loss probability on correlation in the arrival processes.

Since there are two reasons of the loss of type-1 customers: a loss due the lack of additional items and a loss due the competition between type-1 customers, it seems interesting to look at the probabilities $P_1^{\text{item-lack}}$ of a type-1 customer loss due to the lack of additional items and $P_1^{\text{server-busy}}$ of a type-1 customer loss because the server provides service to another type-1 customer separately.

Figure 9 illustrates the dependence of the loss probabilities $P_1^{\text{item-lack}}$ and $P_1^{\text{server-busy}}$ on the threshold N. It is seen from Fig. 9 that, under the fixed values of the system parameters, the main reason of a type-1 customer loss is the lack of additional items. The probability $P_1^{\text{item-lack}}$ decreases when N increases due to the reason explained above. $P_1^{\text{server-busy}}$ increases when N grows. This is clear because fewer type-1 customers are rejected at the entrance to the system with growth of N and the chance for a type-1 customer to meet the server occupied by another type-1 customer increases. It is interesting to note that the worst combination of arrival flows for $P_1^{\text{server-busy}}$ is MMAP$^{0.4}$+MAP^0 while for other performance measures, as a rule, the worst combination was MMAP$^{0.4}$+MAP $^{0.4}$. This is obvious because the zero correlation in the additional items' arrival process implies a more uniform arrival of additional items, relatively low loss of additional items and good conditions for service of type-1 customers. But high correlation in arrivals of type-1 customers essentially deteriorates this favorable situation due to an irregularity in arrivals and, consequently, competition between type-1 customers.

Figure 10 illustrates the dependence of the probability P^{imp} that an arbitrary type-2 customer leaves the system due to impatience and the probability P^{term} that service of an arbitrary type-2 customer is terminated by an arriving type-1 customer on the threshold N.

It is evidently seen from these figures that the probability P^{imp} drastically changes depending on correlation in the arrival processes. For combination of two flows with zero correlation, this probability is less than 0.02 for small N while it is more than 0.5 for combination MMAP$^{0.4}$+MAP $^{0.4}$. Dependencies of the probability of forced termination P^{term} on the threshold N look quite unpredictable. First of all, in contrast to other loss probabilities, the worst case (the maximal loss probability) is achieved here for a combination of two flows with zero correlation, not for correlated flows. This fact can be explained taking into account the curves for P^{imp}. For a combination of two flows with zero correlation, the probability P^{imp} is small. This implies that most customers are not lost during the waiting time in the buffer and start service. Therefore, the probability of service termination is high. The second interesting feature of the curves for P^{term} is that this probability grows when N increases from 1 to 7 and it becomes smaller when N increases from 7 to 19. This feature is explained as follows. As evidenced in Fig. 8, the probability of type-1 customer loss is high for small values of N. Thus, these customers rarely terminate service of type-2 customers. When N grows, the probability P_1^{loss} decreases and the probability P^{term} grows. However, when N becomes large ($N > 7$ in this example), the probability P^{imp} essentially increases, a lot of type-2 customers leave the system

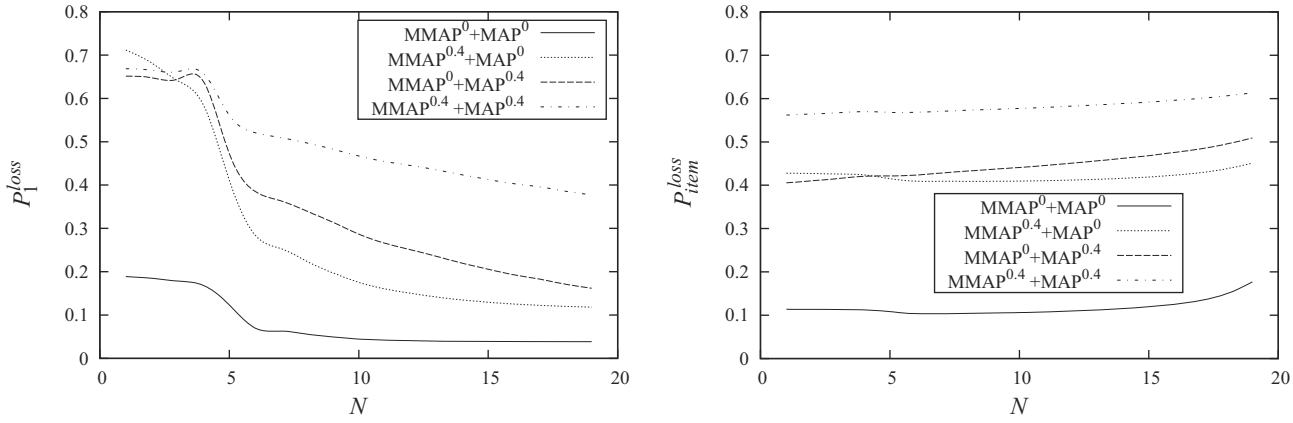


Fig. 8. Dependence of the loss probabilities P_1^{loss} and $P_{\text{item}}^{\text{loss}}$ on the threshold N.

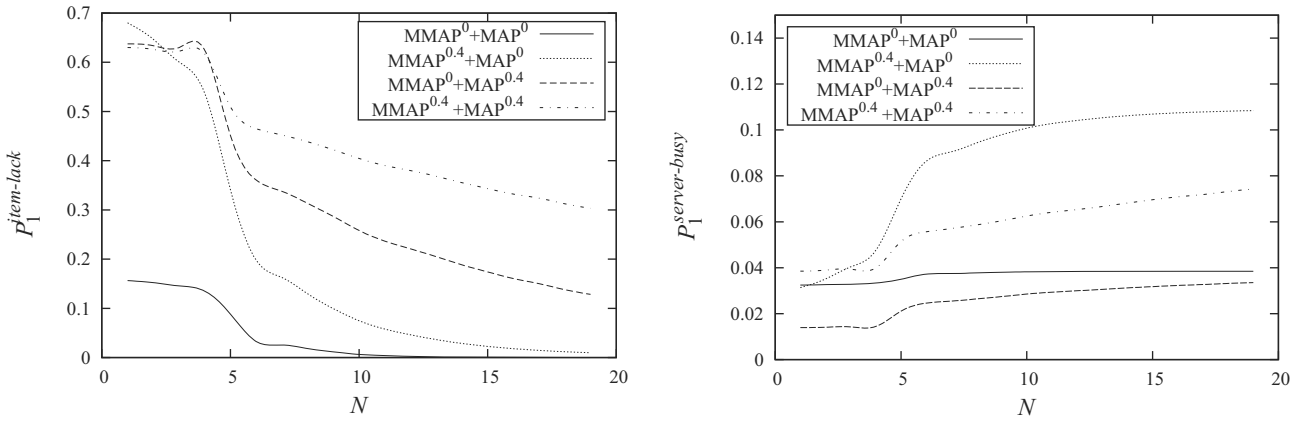


Fig. 9. Dependence of the loss probabilities $P_1^{\text{item-lack}}$ and $P_1^{\text{server-busy}}$ on the threshold N.

without service, and therefore, the termination probability decreases.

Figure 11 illustrates the dependence of the average number $N_{\text{customers}}^{\text{buffer}}$ of customers in the buffer and the average number $N_{\text{item}}^{\text{stock}}$ of additional items in the stock on the threshold N. Figure 12 shows the dependence of the average waiting time V^{wait} of an arbitrary type-2 customer and the average waiting time $V^{\text{wait-patient}}$ of an arbitrary type-2 customer that does not leave the system due to impatience in the system on the threshold N.

It is again seen from Figs. 11 and 12 that the performance measures $N_{\text{customers}}^{\text{buffer}}$, $N_{\text{item}}^{\text{stock}}$, V^{wait} and $V^{\text{wait-patient}}$ strongly depend on correlation in the arrival processes.

It is worth noticing that the results of a large number of numerical experiments show that famous Little's formula holds for the system under study, i.e.,

$$V^{\text{wait}} = {\lambda_2}^{-1} N_{\text{customers}}^{\text{buffer}} \tag{10}$$

as well as the formula

$$V^{\text{wait}} = \alpha^{-1} P^{\text{imp}}. \tag{11}$$

The formula (11) can be derived from (10) as follows. Let us analyse the fraction $N_{\text{customers}}^{\text{buffer}}\alpha/\lambda_2$. The value $N_{\text{customers}}^{\text{buffer}}\alpha$ defines the intensity of type-2 customers' leaving the buffer due to impatience. Thus, this fraction defines the probability of an arbitrary type-2 customer's loss due to impatience, i.e.,

$$P^{\text{imp}} = \frac{N_{\text{customers}}^{\text{buffer}}\alpha}{\lambda_2}.$$

The formula (11) immediately follows from this formula and (10).

7. Conclusion

We analysed the priority queueing system with two types of customers having different requirements on the service time and the number of additional items which should be spent on service. Items arrive at the system at random moments and are accumulated in the stock of a finite capacity. Service of customers is suspended if the stock does not contain a sufficient number of items. To provide better conditions for priority customers, we proposed a

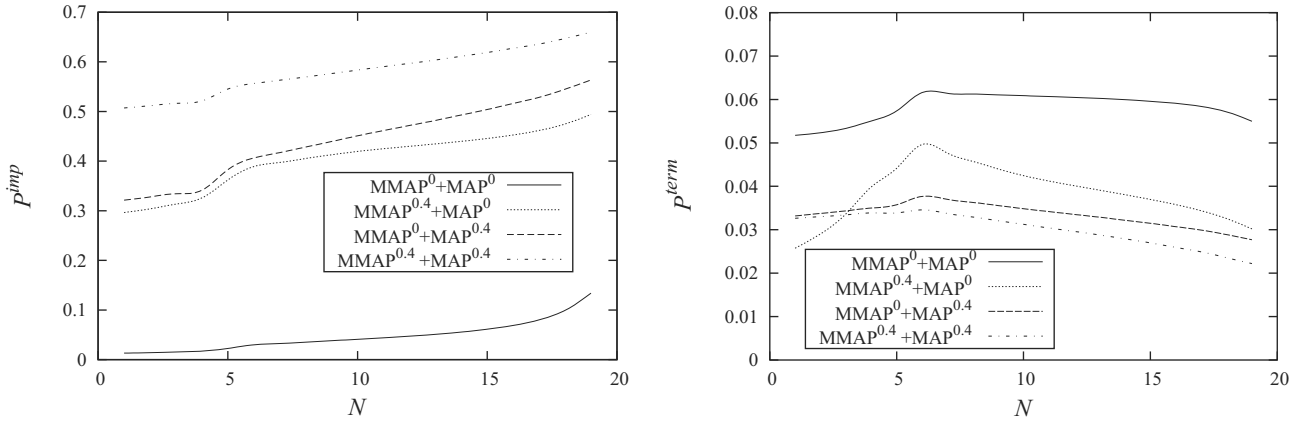


Fig. 10. Dependence of the loss probabilities P^{imp} and P^{term} on the threshold N.

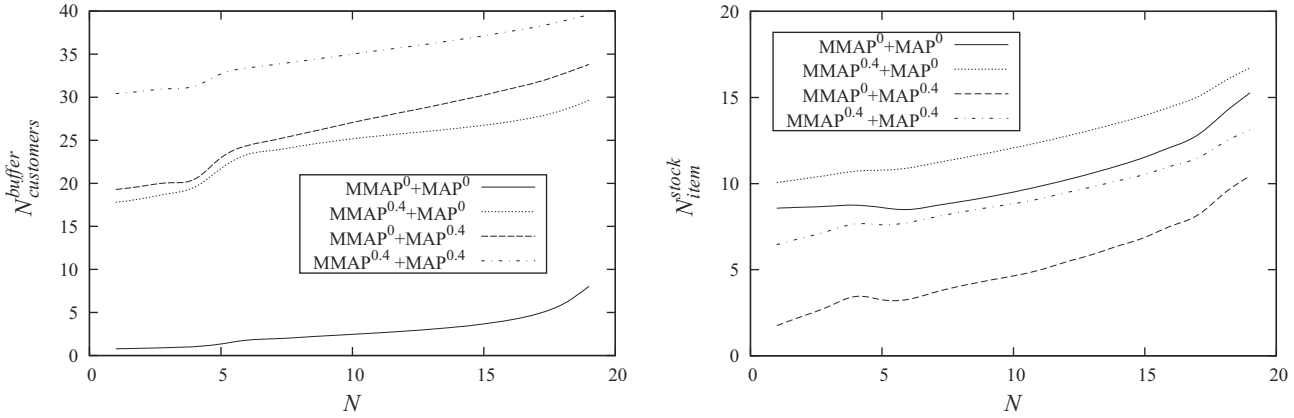


Fig. 11. Dependence of the average number $N^{\text{buffer}}_{\text{customers}}$ of customers in the buffer and average number $N^{\text{stock}}_{\text{item}}$ of items in the stock on the threshold N.

thresholding strategy of access of non-priority customers to the server. Under a fixed value of the threshold, N, we computed the steady state distribution of the state inhomogeneous Markov process describing the system dynamics, the distribution of the waiting time and derived expressions for various performance measures of the system.

Results of numerical experiment, are presented. They show the influence of the threshold on the value of the most important performance measures of the system. Having the presented dependencies, numerous optimization problems can be formulated and solved, e.g., the problem of choosing the value of the threshold providing the minimal waiting time of a non-priority customer conditioned on the fact that the loss probability of the priority customer does not exceed a predefined value. Results of numerical experiments show that the dependence of performance measures on the threshold N is quite complicated, especially when the number of additional items, which should be spent on service of a priority customer, is in the neighborhood of N.

It is difficult to offer some "rule of thumb" for solution of optimization problems. This makes the obtained results important, which allows us to easily and exactly solve various optimization problems. Results of numerical experiments also demonstrate that the ignorance of correlation in the arrival processes of customers and items, which takes place if these processes are assumed to be stationary Poisson, may lead to huge errors in the performance evaluation of the system. It is worth noticing that we analysed the behavior of the system under a fixed arrival flow of additional items and the stock capacity.

However, the presented results can be used in an evident way also for solving the corresponding inventory problem. How large should the capacity of the stock and the speed of additional items delivering be? The results are planned to be extended to the case of the discrete time system similar to the one by Atencia (2014), the batch arrivals and more complicated strategies of control as in the work Gaidamaka *et al.* (2014), the system operating in the random environment (Kim *et al.*, 2014).

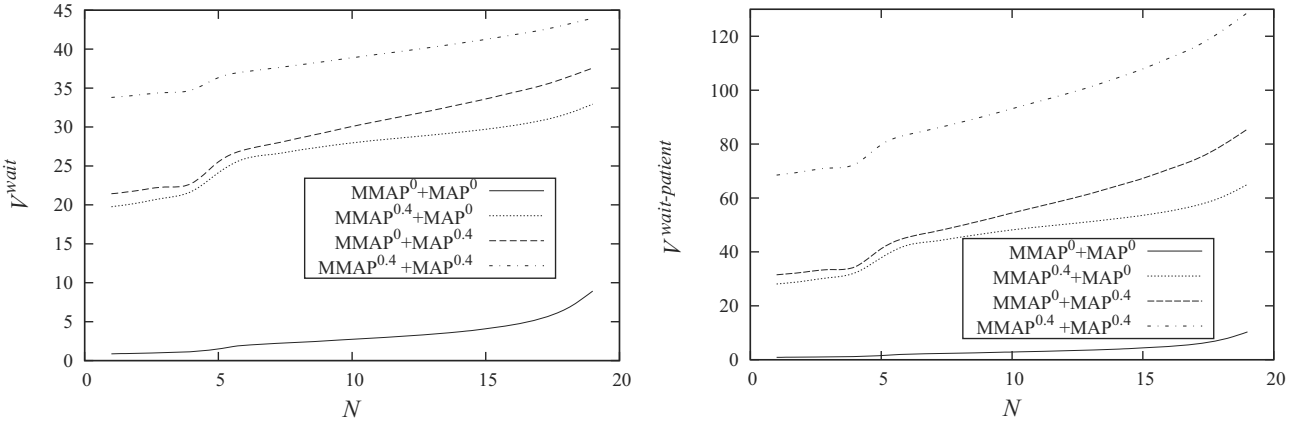


Fig. 12. Dependence of the average waiting times V^{wait} and $V^{\text{wait-patient}}$ on the threshold N.

Acknowledgment

This research was supported by a basic science research program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B01014615), the Ministry of Education and Science of the Russian Federation (the agreement number 02.a03.21.0008 of 24 June 2016), and a basic science research program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant no. 2014R1A1A4A01007517).

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.

Cardoen, B., Demeulemeester, E. and Beliën, J. (2010). Operating room planning and scheduling: A literature review, *European Journal of Operational Research* **201**(3): 921–932.

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

Dantzig, D.v. (1955). Chaînes de Markof dans les ensembles abstraits et applications aux processus avec régions absorbantes et au problème des boucles, *Annales de l'Institut Henri Poincaré* **14**(3): 145–199.

Dudin, A. and Klimenok, V. (1996). Queueing systems with passive servers, *Journal of Applied Mathematics and Stochastic Analysis* **9**(2): 185–204.

Dudin, A., Lee, M. 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.

Dudina, O., Kim, C. and Dudin, S. (2013). Retrial queuing system with Markovian arrival flow and phase-type service time distribution, *Computers & Industrial Engineering* **66**(2): 360–373.

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.

Gelenbe, E. (2015). Synchronising energy harvesting and data packets in a wireless sensor, *Energies* **8**(1): 356–369.

He, Q.-M. (1996). Queues with marked customers, *Advances in Applied Probability* **28**(2) : 567–587.

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

Kim, C., A., D., Dudin, S. and Klimenok, V. (2012). Queueing system with batch arrival of customers in sessions, *Computers and Industrial Engineering* **62**(4): 890–897.

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.

Kim, C., Dudin, S. and Klimenok, V. (2009). The $map/ph/1/n$ queue with flows of customers as model for traffic control in telecommunication networks, *Performance Evaluation* **66**(9): 564–579.

Klimenok, V. and Dudin, A. (2006). Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory, *Queueing Systems* **54**(4): 245–259.

Krishnamoorthy, A., Benny, B. and Shajin, D. (2016a). A revisit to queueing-inventory system with reservation, cancellation and common life time, *OPSEARCH* **54**(2): 336–350, DOI: 10.1007/s12597-016-0278-1.

Krishnamoorthy, A., Shajin, D. and Lakshmy, B. (2016b). On a queueing-inventory with reservation, cancellation, common life time and retrial, *Annals of Operations Research* **247**(1): 365–389.

Krishnamoorthy, A., Shajin, D. and Lakshmy, B. (2016c). Product form solution for some queueing-inventory supply chain problem, *OPSEARCH* **53**(1): 85–102.

Manzini, R., Heragu, S. and Bozer, Y. (2015). Decision models for the design,optimization and management of warehousing and material handling systems, *International Journal of Production Economics* **170**(C): 711–716.

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

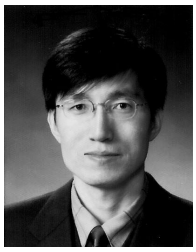
Sharma, V., Mukherji, U., Joseph, V. and Gupta, S. (2010). Optimal energy management policies for energy harvesting sensor nodes, *IEEE Transactions on Wireless Communications* **9**(4): 1326–1336.

Tutuncuoglu, K. and Yener, A. (2012). Optimum transmission policies for battery limited energy harvesting nodes, *IEEE Transactions on Wireless Communications* **11**(3): 1180–1189.

Yang, J. and Ulukus, S. (2012a). Optimal packet scheduling in a multiple access channel with energy harvesting transmitters, *Journal of Communications and Networks* **14**(2): 140–150.

Yang, J. and Ulukus, S. (2012b). Optimal packet scheduling in an energy harvesting communication system, *IEEE Transactions on Communications* **60**(1): 220–230.

Zhao, N. and Lian, Z. (2011). A queueing-inventory system with two classes of customers, *International Journal of Production Economics* **129**(1): 225–231.



Janghyun Baek received the BS, MSc, and PhD degrees in industrial engineering from Seoul National University, Republic of Korea, in 1986, 1988 and 1997, respectively. He worked for Electronics and Telecommunications Research Institute (ETRI), Republic of Korea, from 1989 to 1998. In 1998, he joined the Faculty of College of Engineering, Chonbuk National University, where he is currently a full professor in the Department of Industrial and Information Systems Engineering. His research interests include stochastic processes, optimization theory and their application to communication network. He has published around 50 papers in internationally refereed journals.

Olga S. Dudina graduated from Belarusian State University in 2007. In 2010, she obtained her PhD degree university in probability theory and mathematical statistics from the same university and currently works as a leading scientific researcher in the Research Laboratory of Applied Probabilistic Analysis at BSU. She also works part time at the Peoples' Friendship University of Russia. Her main fields of interests are queueing tandem queueing models with correlated arrival flows, and non-Markovian queueing systems.

Chesoong Kim obtained his MSc and PhD degrees in engineering from the Department of Industrial Engineering at Seoul National University in 1989 and 1993, respectively. He is a full professor and the head of the Department of Industrial Engineering at Sangji University. His research interests are in stochastic processes, queueing theory with particular emphasis on computer and wireless communication networks, queueing networks modeling and their applications. He has published around 60 papers in internationally refereed journals.

Received: 7 September 2016
Revised: 28 November 2016
Re-revised: 21 December 2016
Accepted: 31 January 2017