Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 10(2): 3095–3123.
DOI: 10.3934/math.2025144
Received: 16 January 2025
Revised: 05 February 2025
Accepted: 11 February 2025
Published: 19 February 2025

Research article

Study of a semi-open queueing network with hysteresis control of service regimes

Ciro D'Apice[1], Alexander Dudin[2], Sergei Dudin[2], and Rosanna Manzo[3,*]

[1] Dipartimento di Scienze Aziendali - Management & Innovation Systems, University of Salerno, Via Giovanni Paolo II, 132, Fisciano 84084, Salerno, Italy

[2] Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus

[3] Department of Political and Communication Sciences, University of Salerno, Via Giovanni Paolo II, 132, Fisciano 84084, Salerno, Italy

* **Correspondence:** Email: rmanzo@unisa.it.

Abstract: Semi-open queueing networks are suitable for modeling complex manufacturing, health care, and logistics systems. Such networks are different from more well-known open queueing networks because the number of users, that can be serviced in the network simultaneously is restricted by a finite constant. The network loses customers who arrive when its capacity reaches its limit. This paper examined an analytical model characterized by features like the possibility to capture potential correlations in the arrival process by assuming the marked Markov arrival process and modify service rates in the network's nodes depending on the number of users currently processed in the network. A hysteresis strategy for dynamic service rate selection was assumed. Fixing the thresholds of this strategy, the behavior of the network was determined by a continuous-time multidimensional Markov chain with a finite state that is a quasi-birth-and-death process. An explicit formula for the generator of this process was obtained. Expressions for the computation of network performance measures were derived. Numerical results highlight the dependence of some measures on thresholds defining the control policy, and their use to optimize the system is illustrated.

Keywords: queueing network; hysteresis control; multidimensional Markov chains; modeling
Mathematics Subject Classification: 60K25, 60K30, 60M20

1. Introduction

Semi-open queueing networks (SOQNs), also referred to as open queues with restricted capacity, have garnered the attention of the stochastic modeling research community in recent years; see the survey [1] and the paper [2]. SOQNs are characterized by users, who would like to receive service, arriving from an infinite population of sources and permanently departing from the network after service, as in open queueing networks. However, only a restricted number of users can receive service in the network at the same time, as in closed networks.

A SOQN can be organized into a complex consisting of a set of orders (permissions, windows, threads, tokens, operators, robots, automated guided vehicles, etc.) required to begin service and a core (inner) network, whose nodes are dedicated to providing service for an admitted user. If an arriving user finds an available order in the set, the user enters the core network and sequentially receives service in its nodes. After the end of service in the required sequence of nodes, the user leaves the core network and returns the order, that was received at the entrance to the set of orders. If the set is empty at a user arrival instant, then the user joins a buffer and waits for an order release (this kind of service organization is called back-ordering SOQN, see [3]) or is lost (called user loss SOQN).

As stated in [1], the utilization of SOQNs originates from the study [4] conducted by Avi-Itzhak and Heyman, who examined the efficacy of a multi-programming computer system. In that system, multiple tasks can be executed concurrently, with the central processor managing one task while peripheral devices attend to other tasks.

Brief overviews of the research on SOQN analysis may be found, for example, in [1, 5–7] and more recent publications [3, 8–13] and references therein. Examples of possible applications of SOQNs are listed in [1], as follows: manufacturing material control rules commonly implemented in manufacturing shops to manage the inventory of both finished goods and work-in-process; modeling the operation of vehicle rental providers, where users borrow vehicles from rental depots and return them after the rental time; modeling systems for storage and retrieval based on autonomous vehicles; modeling communication networks based on window flow control; and modeling health-care systems to accurately depict the dynamics of patient-resource waiting, see also [14]. Unmanned manufacturing factories (UMFs) and robotic mobile fulfillment systems (RMFSs) are also very common applications of SOQNs; see, for example, [3, 15–24].

The vast bulk of analytical models of SOQNs currently available in the literature, as well as open and closed queueing networks, assume that the arrival flow of users is determined by the stationary Poisson process. However, it is widely acknowledged that modern real-world arrival processes are poorly fitted by such process. It may be too optimistic to predict real-world systems' performance metrics assuming that the stationary Poisson process describes the arrival flow. This is due to its constant arrival rate, relatively low variance, and zero coefficient of correlation for the sequential inter-arrival periods. Therefore, any irregularity in the arrival process, which is the inherent feature of the majority of real flows, is ignored. Namely, the existence of periods of peak rates leads to congestion and worsens the performance characteristics of the system.

As a much more adequate model of real-world arrival processes, M. Neuts [25] introduced the versatile arrival process, later called the Markov arrival process (MAP), see [26]. More information about the MAP and its extension, such as the batch Markov arrival process (BMAP), can be found, e.g., in [27–31]. To our knowledge, a SOQN with the MAP is only taken into consideration in a small number of articles; see the corresponding references in [7–9].

The MAP assumes that arriving users are of the same type. A more general model of the arrival process is the MMAP (marked Markov arrival process), in which arriving users are heterogeneous; see, e.g., [32]. A SOQN with multi-server nodes and the MMAP, where the type of user predefines the node in the network where the first service of the user is performed, was recently considered in [10].

In this paper, we analyze a SOQN with the MMAP and single-server nodes. The novelty of the model, compared to all existing literature that examines SOQN, consists of the possibility of changing the rate of service in the nodes of the network depending on the number of users admitted to the network.

The literature devoted to the analysis of various controlled queues, including control by the arrival rate, service rate, or number of active servers, is huge. This is easily explained by the fast technological development and consequent possibilities to monitor the state of a real-world system or network and dynamically change the service regime correspondingly. Such control can provide a significant economic effect due to the possibility of regime management in such a way that fast but expensive service regimes are temporarily used only in the case of high system congestion.

Control by the service rate in the nodes can be implemented via the use of equipment with distinct performance (e.g., leasing of channels or routers with different bandwidths in telecommunication networks) or the use of various numbers of workers providing service in a node. High-performance equipment and (or) more staff can be used when the quantity of users receiving service in the network is large, which may imply a long delay of the user in the network, possible dissatisfaction with the quality of service, and possible choice of another service provider in the future. When the network is less congested, it makes sense to use lower-performance equipment and (or) less staff to spend less money on service provision.

As early papers that examine queues with controlled service rates, we can mention [33, 34]. When the system has several available service regimes listed in ascending order of service rate (and, correspondingly, increased cost), it is proved that the optimal control policy is monotone, i.e., of the multi-threshold type. This policy is defined by the set of integer numbers (thresholds). Selection of the service rate at decision moments is performed based on the relationship between the system's current user count and the predetermined thresholds. An application of such a type of policy to the BMAP/G/1 queue with a controlled service time, which can be varied at moments of service completion, was implemented in [35]. The BMAP/SM/1-type queueing system with a multi-threshold control by the service rate, semi-Markov service process, and user retrials in the case of a busy server was analyzed in [36].

One drawback of the threshold policy is that when the current number of users in the system approaches a certain threshold, there is a chance that the service rate will fluctuate often, a phenomenon known as oscillation. To overcome this negative phenomenon (due to the possible time loss or charge for the switching), the so-called hysteresis strategy has been offered. This type of strategy assumes that each policy threshold can be split into two thresholds. The rate of service increases when the larger of these two thresholds is exceeded. When the length of a queue drops to a smaller threshold, the service slows. The system stays in the prior regime when the queue length falls between the thresholds. This delay in switching (called hysteresis in Greek) leads to a decrease in the frequency of regime switching.

The M^x/G/1-type queueing system with two available service regimes, a batch stationary Poisson arrival process, a general distribution of service time, an account of switch-over times, and a hysteresis policy of control was considered in [37, 38]. The result was generalized to the BMAP/G/1-type queueing system in [39]. An analysis of the queue with hysteresis control is more tricky than that of the same type of queue with threshold control. This is because, given the fixed values of the thresholds of the threshold policy, the choice of a service rate at a decision instant is uniquely determined by the current number of users in the network. In the case of the hysteresis policy of control, selection depends also on the service rate used before the moment of decision-making.

In this paper, we consider a user loss SOQN with arbitrary topology, an arriving flow defined by the MMAP, the availability of a finite number $L,\ L \geq 2$, of service regimes, stochastic routing, and the hysteresis type of control by the service rates in the nodes. An exact algorithmic study of the steady-state dynamic of the network is implemented and numerically illustrated.

The structure of the paper is as follows. The description of the SOQN under study is presented in Section 2. The topology of the network is explained, a necessary notation is presented, and the hysteresis-type control policy is defined in detail. In Section 3, a multidimensional continuous-time Markov chain defining the system behavior under the fixed set of thresholds of the control strategy is introduced. The methodology tracing back to [40] is used here for keeping track of the number of users residing in all nodes of the network. The explicit form of the infinitesimal generator of this Markov chain is derived. Formulas for the calculation of a variety of performance measures of the network via the vectors of the steady-state probabilities of the Markov chain are given in Section 4. A numerical illustration of how the algorithmic results can be used for the analysis of the network consisting of 3 nodes, having 3 different service regimes, and an opportunity to simultaneously accommodate up to 40 users is presented in Section 5. Section 6 summarizes the results of the implemented study and briefly discusses possible directions for further research.

2. An explanation of the model

Let us consider a SOQN that consists of K nodes. The illustration of the network structure is presented in Figure 1.

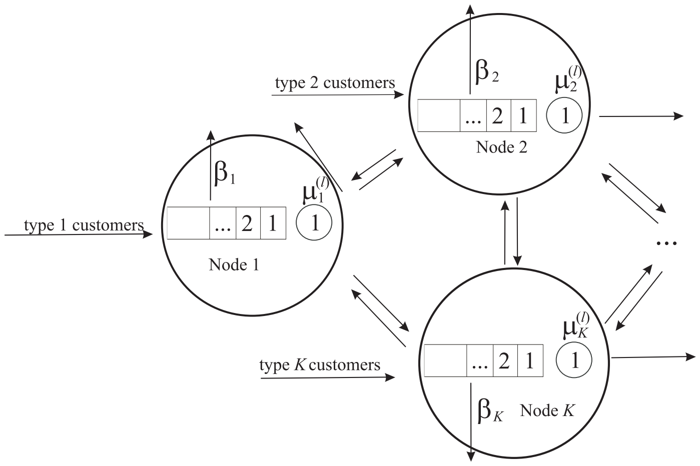

Figure 1. Structure of the network.

Arrivals of users follow a marked Markov arrival process (MMAP). The irreducible underlying Markov chain $\{v_t,\ t \geq 0\}$ with the state space $\{1, \dots, V\}$ governs this process. The generator H of this chain is represented by the form

$$H = \sum_{k=0}^{K} H_k$$

where the diagonal entries of the matrix H_0 are negative. The rates at which the chain $\{\nu_t,\ t \geq 0\}$ exits the appropriate states are determined by their moduli. The non-diagonal entries of the matrix H_0 are nonnegative and specify the rates at which the chain $\{\nu_t,\ t \geq 0\}$ makes transitions without users arriving. The nonnegative entries in the matrix H_k specify the rates at which the chain $\{\nu_t,\ t \geq 0\}$ transits when a user arrives at the network's k-th node, $k = \overline{1,K}$. The expression $k = \overline{1,K}$ indicates that the integer variable k takes values in the set $\{1, 2, \dots, K\}$.

The average rate of arrivals to the k-th node of the network is calculated as $\lambda_k = \boldsymbol{\theta} H_k \mathbf{e},\ k = \overline{1,K}$, where the row vector $\boldsymbol{\theta}$ defines the stationary probability distribution of the states of the chain $\{\nu_t,\ t \geq 0\}$. The vector $\boldsymbol{\theta}$ satisfies the equations $\boldsymbol{\theta} H = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$. Here, $\mathbf{0}$ represents the row vector with 0s, and $\mathbf{e}$ represents the column vector with 1s. The total rate of arrivals is defined as $\lambda = \sum\limits_{k=1}^{K} \lambda_k$.

Various details on the MMAP, its properties, and attributes, such as the distribution of moments of inter-arrival times, the coefficients of variation, and the correlation of inter-arrival times, can be found, for example, in [27–30, 32].

The network can process a maximum of N users simultaneously. When there are N users on the network, the user who arrives is regarded as lost. Every network's node functions as a queueing system having a single-server and a buffer with a capacity that is large enough to ensure that users are never lost.

The network can serve users in L, $L \geq 2$, regimes. If the current service regime is l, the distribution of service time in the k-th node is exponential with intensity $\mu_k^{(l)},\ l = \overline{1,L}$. After service completion at the k-th node, with probability $p_{k,k'}$ the user transits for service to the k'-th node, $k' = \overline{1,K},\ k' \neq k$, or with probability $p_{k,0}$ ends service in the network. Here, $p_{k,k} = 0,\ \sum\limits_{k'=0}^{K} p_{k,k'} = 1$ for all $k,\ k = \overline{1,K}$.

The switching between the available service regimes is implemented by the following control strategy.

Let two sets of integer numbers (thresholds), $\{L_1^+, L_2^+, \dots, L_{L-1}^+\}$ and $\{L_1^-, L_2^-, \dots, L_{L-1}^-\}$, such that

$$0 \leq L_1^- \leq L_1^+ < L_2^- \leq L_2^+ < \cdots < L_{L-1}^- \leq L_{L-1}^+ < N$$

be fixed.

We assume that the network instantly transitions to operating in the $(l+1)$-th regime if the number of users exceeds the threshold L_l^+ and the current regime is l. For example, if the network operates in the first service regime, the number of users in the network is L_1^+, a new user is admitted to the network, and the service is immediately switched to the second regime.

If some user leaves the network when it operates in the l-th regime, $2 \leq l \leq L$, and the number of users in the network drops to the value L_{l-1}^-, the network immediately switches to operation in the $(l-1)$-th regime.

Let us present the sets listed below:

$$W_1 = \{0, 1, \dots, L_1^-\},\ W_l = \{L_{l-1}^+ + 1, L_{l-1}^+ + 2, \dots, L_l^-\},\ l = \overline{2, L-1},$$

$$W_L = \{L_{L-1}^+ + 1, L_{L-1}^+ + 2, \dots, N\},$$

$$F_l = \{L_l^- + 1, L_l^- + 2, \dots, L_l^+\},\ l = \overline{1, L-1}.$$

The network functions in the l-th regime if the quantity of users in the network equals n and $n \in W_l$, $l = \overline{1, L}$, as stated in the model description. If $n \in F_l$, $l = \overline{1, L-1}$, the network can operate in both the l-th and $(l+1)$-th regime, and the used regime has to specified.

The impatience of users (leaving the queue without receiving service because they waited in the buffer for too long) is an essential feature of many real-world systems. For a survey of the literature dealing with queues with impatient users, see, e.g., [41]. To take the impatience phenomenon into account, we assume the impatience of the users waiting in the buffers of the nodes of the network. A user in the k-th buffer reneges after an amount of time that is exponentially distributed with intensity β_k, $k = \overline{1, K}$, independently from other users who are waiting. Such a user permanently reneges from the network (is lost).

We aim to analyze the invariant distribution of the network states and derive expressions for the calculation of the main performance characteristics of the network under the arbitrarily fixed values of the thresholds. As a result, a variety of optimization issues can be formulated and resolved.

3. The process that describes the network states' dynamics

The continuous-time Markov chain

$$\zeta_t = \{n_t, i_t,\ v_t, m_t^{(1)}, \dots, m_t^{(K)}\},\ t \geq 0,$$

can be used to define the behavior of the considered network.

Here, during the instant t,

- the component n_t defines the quantity of users in the network, $n_t = \overline{0, N}$;
- the component i_t is an indicator of the current regime of the network operation. It is defined only for the values n_t such that $n_t \in F_l$, $l = \overline{1, L-1}$. It admits values 0 or 1. Namely, the value $i_t = 0$ indicates the operation in the l-th regime, and $i_t = 1$ indicates the operation in the $(l+1)$-th regime;
- the component v_t specifies the status of the underlying process of the MMAP, $v_t = \overline{1, V}$;
- the component $m_t^{(k)}$ defines the quantity of users in the k-th node, $m_t^{(k)} = \overline{0, n_t}$, $\sum\limits_{k=1}^{K} m_t^{(k)} = n_t$, $k = \overline{1, K}$.

Let us call the set $\{n_t, i_t,\ v_t, m_t^{(1)}, \dots, m_t^{(K)}\}$ of the states of Markov chain ζ_t, $t \geq 0$ enumerated in the reverse lexicographic order of the processes $m_t^{(1)}, \dots, m_t^{(K)}$ and the direct lexicographic order of the processes (i_t, v_t) as level n_t.

The Markov chain $\{\zeta_t,\ t \geq 0\}$ is regular and irreducible. Its state space is finite. Therefore, the invariant probabilities of the states of this chain

$$\pi(n, i, v, m^{(1)}, \dots, m^{(K)}) = \lim_{t\to\infty} P\{n_t = n, i_t = i, v_t = v, m_t^{(1)} = m^{(1)}, \dots, m_t^{(K)} = m^{(K)}\}$$

exist for all possible values of the queuing network's parameters.

The probability of the states that correspond to the level n, $n = \overline{0, N}$, can be formed as row vectors $\boldsymbol{\pi}_n$. The probability vectors $\boldsymbol{\pi}_n$, $n = \overline{0, N}$, can be computed as the solution of the system of linear algebraic equations (known as Chapman-Kolmogorov, balance, or equilibrium equations):

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_N)Q = \mathbf{0},\quad (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_N)\mathbf{e} = 1 \tag{1}$$

where the matrix Q represents the generator of the Markov chain $\{\zeta_t,\ t \geq 0\}$.

To calculate the vectors $\boldsymbol{\pi}_n$, $n = \overline{0,N}$, the exact expression for the generator Q must be obtained.

Let us present the following notation for future use in this paper:

- O is a zero matrix of the suitable size, and I is the identity matrix. If the size is not clear from context, it is indicated by a suffix, e.g., I_K is the identity matrix of size K;
- the symbols $\otimes$ and $\oplus$ represent the Kronecker product and sum of matrices, respectively; refer to [42];
- $\mathbf{b}^{(k)}$ is a vector of size K defined as $\mathbf{b}^{(k)} = \{\underbrace{0,0,\ldots,0,1}_{k},0,\ldots,0\},\ k = \overline{1,K}$;
- diag$\{\ldots\}$ denotes the diagonal matrix with the diagonal entries shown in brackets;
- $\Omega^{(l)}$, $l = \overline{1,L}$, is the square matrix of dimension K defined as:

$$\Omega^{(l)} = \text{diag}\{\mu_k^{(l)},\ k = \overline{1,K}\}(-I + P), l = \overline{1,L},$$

  where the matrix P with the entries $p_{k,k'}$, $k, k' = \overline{1,K}$, defines the one-step transition probabilities of a user within the network;
- $\mathbf{p}^{(l)}$, $l = \overline{1,L}$, is the column vector defined as $\mathbf{p}^{(l)} = (p_1^{(l)}, p_2^{(l)}, \ldots, p_K^{(l)})^T$ where $p_k^{(l)} = p_{k,0}\mu_k^{(l)},\ k =\overline{1,K}$;
- $\boldsymbol{\beta}$ is the column vector defined as $\boldsymbol{\beta} = (\beta_1,\beta_2,\ldots,\beta_K)^T$.

Let us first introduce the set of matrices that describe the behaviour of the K-dimensional stochastic process $\mathbf{m}_t = \{m_t^{(1)},\ldots,m_t^{(K)}\},\ t \geq 0$, defining the quantity of users in each node of the network. To define the transition rates of this process, conditioned on the knowledge that n users are present in the network working in the l-th regime at instant t, we need the following matrices:

a) the matrix $\mathcal{T}_n(\Omega^{(l)})$ establishes the process $\mathbf{m}_t,\ t \geq 0$, of transition intensities at the precise instant when a user ends the service in one network's node and moves on to another node, $n = \overline{1,N}$, $l = \overline{1,L}$.
   The following six steps make up the procedure for calculating the matrices $\mathcal{T}_n(\Omega^{(l)}),\ n = \overline{1,N},\ l = \overline{1,L}$:
   (1) Determine the matrices $\Omega_k^{(l)}$, $k = \overline{1,K-2}$, that result from subtracting the $K-2-k$ first rows and columns from the matrix $\Omega^{(l)}$.
   (2) Calculate the set of matrices $X_{n,k}^{(r,l)}$ using the recursive formulas:

$$X_{n,k}^{(0,l)} = \omega_{r_k^{(l)},1}^{k,l}, n = \overline{1,N},\ k = \overline{1,K-2},$$

$$X_{n,k}^{(r,l)} = \begin{pmatrix} \omega_{r_k^{(l)}-r,1}^{k,l} I & O & \cdots & O \\ X_{1,k}^{(r-1,l)} & \omega_{r_k^{(l)}-r,1}^{k,l} I & \cdots & O \\ O & X_{2,k}^{(r-1,l)} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \omega_{r_k^{(l)}-r,1}^{k,l} I \\ O & O & \cdots & X_{n,k}^{(r-1,l)} \end{pmatrix},$$

$$r=\overline{1,r_k^{(l)}-2},\ n=\overline{1,N},\ k=\overline{1,K-2},\ l=\overline{1,L},$$

where $\omega^{k,l}_{i_1,i_2}$ is the (i_1,i_2)th entry of the matrix $\Omega^{(l)}_k$, and $r^{(l)}_k$ represents how many rows there are in the matrix $\Omega^{(l)}_k$.

(3) Utilizing the recursive formulas, determine the set of the matrices $Y^{(r,l)}_{n,k}$:

$$Y^{(0,l)}_{n,k}=\omega^{k,l}_{1,r^{(l)}_k},\ n=\overline{0,N-1},\ k=\overline{1,K-2},$$

$$Y^{(r,l)}_{n,k}=\begin{pmatrix} \omega^{k,l}_{1,r^{(l)}_k-r}I & Y^{(r-1,l)}_{0,k} & O & \cdots & O & O\\ O & \omega^{k,l}_{1,r^{(l)}_k-r}I & Y^{(r-1,l)}_{1,k} & \cdots & O & O\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ O & O & O & \cdots & \omega^{k,l}_{1,r^{(l)}_k-r}I & Y^{(r-1,l)}_{n,k} \end{pmatrix},$$

$$r=\overline{1,r^{(l)}_k-2},\ n=\overline{0,N-1},\ k=\overline{1,K-2},\ l=\overline{1,L}.$$

(4) Calculate the matrices $X^{(l)}_{n,k}=X^{(r_k-2,l)}_{n,k}$, $n=\overline{1,N}$, and $Y^{(l)}_{n,k}=Y^{(r_k-2,l)}_{n,k}$, $n=\overline{0,N-1}$, $k=\overline{1,K-2}$, $l=\overline{1,L}$.

(5) Calculate the matrices $\mathcal{T}^{(k,l)}_n=\mathcal{T}^{(k,l)}_n(\Omega^{(l)})$ using the recursive formulas:

$$\mathcal{T}^{(0,l)}_n=\begin{pmatrix} O & \Omega^{(l)}_{K-1,K} & O & \cdots & O & O\\ \Omega^{(l)}_{K,K-1} & O & \Omega^{(l)}_{K-1,K} & \cdots & O & O\\ O & \Omega^{(l)}_{K,K-1} & O & \cdots & O & O\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ O & O & O & \cdots & O & \Omega^{(l)}_{K-1,K}\\ O & O & O & \cdots & \Omega^{(l)}_{K,K-1} & O \end{pmatrix},$$

$$n=\overline{1,N},\ l=\overline{1,L},$$

$$\mathcal{T}^{(k,l)}_n=\begin{pmatrix} O & Y^{(l)}_{0,k} & O & \cdots & O & O\\ X^{(l)}_{1,k} & \mathcal{T}^{(k-1,l)}_1 & Y^{(l)}_{1,k} & \cdots & O & O\\ O & X^{(l)}_{2,k} & \mathcal{T}^{(k-1,l)}_2 & \cdots & O & O\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ O & O & O & \cdots & \mathcal{T}^{(k-1,l)}_{n-1} & Y^{(l)}_{n-1,k}\\ O & O & O & \cdots & X^{(l)}_{n,k} & \mathcal{T}^{(k-1,l)}_n \end{pmatrix},$$

$$n=\overline{1,N},\ k=\overline{1,K-2},\ l=\overline{1,L}.$$

(6) Calculate the matrices $\mathcal{T}_n(\Omega^{(l)})$ as

$$\mathcal{T}_n(\Omega^{(l)})=\mathcal{T}^{(K-2,l)}_n,\ n=\overline{1,N}.$$

b) The matrix $\mathcal{S}_n(\mathbf{p}^{(l)})$ defines the intensities of the process $\mathbf{m}_t$, $t\geq 0$, of transitions that occur at the moment of a user service completion in some node and its departure from the network, $n=\overline{1,N}$, $l=\overline{1,L}$.

The matrices $\mathcal{S}_n(\mathbf{p}^{(l)})$, $n=\overline{1,N}$, $l=\overline{1,L}$, can be found as

$$\mathcal{S}_n(\mathbf{p}^{(l)})=\mathcal{S}_n^{(K-1,l)}(\mathbf{p}^{(l)}),\ n=\overline{1,N},\ l=\overline{1,L},$$

where the matrices $\mathcal{S}_n^{(K-1,l)}(\mathbf{p}^{(l)})$ are recursively computed as:

$$\mathcal{S}_n^{(0,l)}(\mathbf{p}^{(l)})=\mathbf{p}_K^{(l)},$$

$$\mathcal{S}_n^{(k,l)}(\mathbf{p}^{(l)})=\begin{pmatrix} \mathbf{p}_{K-k}^{(l)}I & O & \cdots & O \\ \mathcal{S}_1^{(k-1,l)}(\mathbf{p}^{(l)}) & \mathbf{p}_{K-k}^{(l)}I & \cdots & O \\ O & \mathcal{S}_2^{(k-1,l)}(\mathbf{p}^{(l)}) & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \mathbf{p}_{K-k}^{(l)}I \\ O & O & \cdots & \mathcal{S}_n^{(k-1,l)}(\mathbf{p}^{(l)}) \end{pmatrix},$$

$$k=\overline{1,K-1},\ n=\overline{1,N},\ l=\overline{1,L}.$$

c) The matrix $\mathcal{I}_n(\boldsymbol{\beta})$ defines the intensities of the process $\mathbf{m}_t$, $t\geq 0$, of transitions that happen at the moment of some user loss due to impatience, $n=\overline{2,N}$. The matrices $\mathcal{I}_n(\boldsymbol{\beta})$, $n=\overline{2,N}$, can be found as $\mathcal{I}_n(\boldsymbol{\beta})=\mathcal{I}_n^{(K-1)}(\boldsymbol{\beta})$, $n=\overline{2,N}$, where the matrices $\mathcal{I}_n^{(K-1)}(\boldsymbol{\beta})$, $n=\overline{2,N}$, are recursively obtained as

$$\mathcal{I}_n^{(0)}(\boldsymbol{\beta})=\max\{0,n-1\}\beta_K,$$

$$\mathcal{I}_n^{(k)}(\boldsymbol{\beta})=\begin{pmatrix} \max\{0,n-1\}\beta_{K-k}I & O & \cdots & O \\ \mathcal{I}_1^{(k-1)}(\boldsymbol{\beta}) & \max\{0,n-2\}\beta_{K-k}I & \cdots & O \\ O & \mathcal{I}_2^{(k-1)}(\boldsymbol{\beta}) & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \max\{0,0\}\beta_{K-k}I \\ O & O & \cdots & \mathcal{I}_n^{(k-1)}(\boldsymbol{\beta}) \end{pmatrix},$$

$$k=\overline{1,K-1},\ n=\overline{2,N}.$$

Hereinafter, we assume that $\mathcal{I}_1(\boldsymbol{\beta})$ is a zero matrix.

d) The probabilities of the process $\mathbf{m}_t$ transition at the time of a type-k user admittance to the network are defined by the matrix $\mathcal{P}_n(\mathbf{b}^{(k)})$, $k=\overline{1,K}$, $n=\overline{0,N-1}$. The paper [43] contains the algorithm for computing matrices $\mathcal{P}_n(\mathbf{b}^{(r)})$. Using the denotations of our paper, we represent that algorithm below for the benefit of the reader. The matrices $\mathcal{P}_n(\mathbf{b})$, $n=\overline{0,N-1}$, where $\mathbf{b}=(b_1,\ldots,b_K)$, $\mathbf{b}\in\{\mathbf{b}^{(1)},\mathbf{b}^{(2)},\ldots,\mathbf{b}^{(K)}\}$, can be computed as

$$\mathcal{P}_0(\mathbf{b})=\mathbf{b},\ \mathcal{P}_n(\mathbf{b})=\mathcal{P}_n^{(K-2)},\ n=\overline{1,N-1},$$

where the matrices $\mathcal{P}_n^{(k)}$ of size $(n+1)\times(n+2)$, $n=\overline{1,N-1}$, are defined as

$$\mathcal{P}_n^{(0)}=\begin{pmatrix} b_{K-1} & b_K & 0 & \cdots & 0 & 0 \\ 0 & b_{K-1} & b_K & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_{K-1} & b_K \end{pmatrix},$$

$$\mathcal{P}_n^{(k)} = \begin{pmatrix} \mathbf{b}_{K-k-1} & \tilde{\mathbf{b}}^{(k)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{b}_{K-k-1}I & \mathcal{P}_1^{(k-1)} & O & \cdots & O & O \\ \mathbf{0}^T & O & \mathbf{b}_{K-k-1}I & \mathcal{P}_2^{(k-1)} & \cdots & O & O \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}^T & O & O & O & \cdots & \mathbf{b}_{K-k-1}I & \mathcal{P}_n^{(k-1)} \end{pmatrix}, k = \overline{1, K-2},$$

Here, the vectors $\tilde{\mathbf{b}}^{(k)}$ are defined as

$$\tilde{\mathbf{b}}^{(k)} = (b_{K-k}, b_{K-k+1}, \dots, b_K),\ k = \overline{1, K-2}.$$

e) The overall rate at which the process $\mathbf{m}_t$ exits the respective state is given by the moduli of the diagonal components of the diagonal matrix $\mathcal{E}_n^{(l)}$, $n = \overline{1, N}$, $l = \overline{1, L}$. The matrices $\mathcal{E}_n^{(l)}$ are given by the formula

$$\mathcal{E}_n^{(l)} = -\mathrm{diag}\{\mathcal{T}_n(\Omega^{(l)})\mathbf{e} + \mathcal{S}_n(\mathbf{p}^{(l)})\mathbf{e} + \mathcal{I}_n(\boldsymbol{\beta})\mathbf{e}\},\ n = \overline{1, N}.$$

The presented formulas and algorithms used for the description of transition rates and probabilities of the process $\mathbf{m}_t$ are based on the ideas presented in [40] for the simultaneous description of the phase-type service processes in several independent servers.

Remark 1. *A fact worth mentioning is that here, we compute the superfluous number of matrices* $\mathcal{T}_n^{(l)}(\Omega^{(l)})$, $\mathcal{E}_n^{(l)}$, *and* $\mathcal{S}_n^{(l)}(\mathbf{p}^{(l)})$ *because if the network operates in the l-th regime, then the number n of users in the network should admit not an arbitrary value in the range from 0 to N but should belong to the interval* $[0, L_1^+]$, *if* $l = 1$, $[L_{l-1}^- + 1, L_l^+]$, *if* $l = \overline{2, L-1}$, *and* $[L_{L-1}^- + 1, N]$, *if* $l = L$. *However, because we intend to build up the surfaces illustrating certain dependencies of the network performance indicators on the thresholds and solve the optimization problem, we would like to compute the generator Q of the Markov chain* ζ_t, $t \geq 0$, *for all possible sets of the thresholds* L_l^+ *and* L_l^-, $l = \overline{1, L-1}$. *Therefore, if we compute all the matrices* $\mathcal{T}_n^{(l)}(\Omega^{(l)})$, $\mathcal{E}_n^{(l)}$, *and* $\mathcal{S}_n^{(l)}(\mathbf{p}^{(l)})$ *for every value of n in the range from 0 to N and all values of l in the range from 1 to L from the early beginning, we will avoid redundant repeated computations during the building of the surfaces and the solution of the optimization problem. If it is required to make calculations only for the fixed set of the thresholds, computation of these matrices can be implemented only for the values on N from the corresponding diapason.*

After calculating the matrices mentioned above, which completely characterize the service process of users $\mathbf{m}_t$, $t \geq 0$, in all nodes of the network, we are ready to formulate the following statement.

Theorem 1. *The infinitesimal generator Q of the Markov chain* ζ_t, $t \geq 0$, *under study has a block-tridiagonal structure.*

The diagonal blocks $Q_{n,n}$, $n = \overline{0, N}$, *of the generator are given by formulas:*

$$Q_{0,0} = H_0,$$

$$Q_{n,n} = H_0 \oplus (\mathcal{T}_n(\Omega^{(l)}) + \mathcal{E}_n^{(l)}),\ \textit{for } n \neq 0,\ n \neq N, n \in W_l,\ l = \overline{1, L},$$

$$Q_{n,n} = \begin{pmatrix} H_0 \oplus (\mathcal{T}_n(\Omega^{(l)}) + \mathcal{E}_n^{(l)}) & O \\ O & H_0 \oplus (\mathcal{T}_n(\Omega^{(l+1)}) + \mathcal{E}_n^{(l+1)}) \end{pmatrix},\ \textit{for } n \in F_l,\ l = \overline{1, L-1},$$

$$Q_{N,N} = H \oplus (\mathcal{T}_N(\Omega^{(L)}) + \mathcal{E}_N^{(L)}).$$

The subdiagonal blocks $Q_{n,n-1}$, $n = \overline{1,N}$, *are given by formulas:*

$$Q_{n,n-1} = \begin{pmatrix} O & I_V \otimes (\mathcal{S}_n(\mathbf{p}^{(l+1)}) + \mathcal{I}_n(\boldsymbol{\beta})) \end{pmatrix}, \text{ for } n = L_l^+ + 1 \neq L_l^- + 1, \ l = \overline{1, L-1}.$$

$$Q_{n,n-1} = I_V \otimes (\mathcal{S}_n(\mathbf{p}^{(l+1)}) + \mathcal{I}_n(\boldsymbol{\beta})), \text{ for } n = L_l^+ + 1 = L_l^- + 1, \ l = \overline{1, L-1}.$$

$$Q_{n,n-1} = I_V \otimes (\mathcal{S}_n(\mathbf{p}^{(l)}) + \mathcal{I}_n(\boldsymbol{\beta})), \ n \in W_l, \ l = \overline{1,L}, \text{ if } n \neq L_l^+ + 1, \ l = \overline{1, L-1}.$$

$$Q_{n,n-1} = \begin{pmatrix} I_V \otimes (\mathcal{S}_n(\mathbf{p}^{(l)}) + \mathcal{I}_n(\boldsymbol{\beta})) & O \\ O & I_V \otimes (\mathcal{S}_n(\mathbf{p}^{(l+1)}) + \mathcal{I}_n(\boldsymbol{\beta})) \end{pmatrix},$$

$$n \in F_l, \text{ for } n \neq L_l^- + 1, \ l = \overline{1, L-1},$$

$$Q_{n,n-1} = \begin{pmatrix} I_V \otimes (\mathcal{S}_n(\mathbf{p}^{(l)}) + \mathcal{I}_n(\boldsymbol{\beta})) \\ I_V \otimes (\mathcal{S}_n(\mathbf{p}^{(l+1)}) + \mathcal{I}_n(\boldsymbol{\beta})) \end{pmatrix}, \text{ for } n = L_l^- + 1 \neq L_l^+ + 1, \ l = \overline{1, L-1}.$$

The updiagonal blocks $Q_{n,n+1}$, $n = \overline{0, N-1}$, *are given by formulas:*

$$Q_{n,n+1} = \sum_{k=1}^{K} H_k \otimes \mathcal{P}_n(\mathbf{b}_k), \ n \in W_l, \text{ for } n \neq L_l^-, \ l = \overline{1,L}, \text{ or } n = L_l^- = L_l^+, \ l = \overline{1, L-1},$$

$$Q_{n,n+1} = \begin{pmatrix} \sum\limits_{k=1}^{K} H_k \otimes \mathcal{P}_n(\mathbf{b}_k) & O \end{pmatrix}, \text{ for } n = L_l^- \neq L_l^+, \ l = \overline{1, L-1},$$

$$Q_{n,n+1} = \begin{pmatrix} \sum\limits_{k=1}^{K} H_k \otimes \mathcal{P}_n(\mathbf{b}_k) & O \\ O & \sum\limits_{k=1}^{K} H_k \otimes \mathcal{P}_n(\mathbf{b}_k) \end{pmatrix}, \text{ for } n \in F_l, \ n \neq L_l^+, \ l = \overline{1, L-1},$$

$$Q_{n,n+1} = \begin{pmatrix} \sum\limits_{k=1}^{K} H_k \otimes \mathcal{P}_n(\mathbf{b}_k) \\ \sum\limits_{k=1}^{K} H_k \otimes \mathcal{P}_n(\mathbf{b}_k) \end{pmatrix}, \text{ for } n = L_l^+ \neq L_l^-, \ l = \overline{1, L-1}.$$

Proof. To implement the theorem's proof, every potential transition of the Markov chain ζ_t over an infinitesimally short interval is analyzed, and the transition intensities are rewritten in block matrix form. The simultaneous transition rates of two independent Markov chains, v_t and $\mathbf{m}_t$, are defined via the Kronecker product of matrices.

The meaning of the blocks of the generator is transparent taking into account the described above probabilistic meaning of the matrices H_0, H_1 and $\mathcal{T}_n^{(l)}(\Omega^{(l)})$, $\mathcal{E}_n^{(l)}$, and $\mathcal{S}_n^{(l)}(\mathbf{p}^{(l)})$. The off-diagonal components of the diagonal blocks $Q_{n,n}$ define the rates of transition of the Markov chain ζ_t inside the level n. The modules of the negative diagonal entries of the blocks $Q_{n,n}$ define the rates of the departure of the Markov chain ζ_t from the states that belong to the level n. The entries of the blocks $Q_{n,n-1}$ define transition intensities of the Markov chain ζ_t from the level n to the level $n-1$. Such transitions can occur during service completion of some user in the network or user loss due to impatience. The entries of the blocks $Q_{n,n+1}$ define transition intensities of the Markov chain ζ_t from the level n to the level $n+1$. Such transitions can occur at the moments of a new user arrival and admission. Different sizes of some blocks are explained by the possibility of transitions of the number n of a level between the sets W_l and F_l, which imply the change (between one and two) of the cardinality of the state space of the component i_t. □

Remark 2. *The number of blocks in the matrix Q depends on the maximum quantity N of users in the network but does not depend on the number L of the available service regimes. But the size of these blocks may depend on L and the relations between the thresholds L_l^+ and L_l^-, $l = \overline{1, L-1}$. This size is minimal when the service rate control strategy is of the threshold, but not the hysteresis, type, i.e., $L_l^+ = L_l^-$, $l = \overline{1, L-1}$. The square matrix Q may have a huge size. Even in the simpler case of the threshold strategy, it is equal to $V \sum\limits_{n=0}^{N} J_n$ where J_m is the quantity of variants to distribute m users among K existing nodes, which is defined by the formula*

$$J_m = \binom{m+K-1}{K-1} = \frac{(m+K-1)!}{m!(K-1)!}, \; m = \overline{1, N}, \; J_0 = 1. \tag{2}$$

Therefore, certain efficient algorithms that use the generator's sparse structure are required to solve the system (1). Specifically, the approach from [44] can be suggested to determine the queueing network's and Markov chain's stationary probability distribution.

4. Performance measures

Following the computation of the probability vectors $\boldsymbol{\pi}_n$, $n = \overline{0, N}$, we may compute various performance metrics of the queuing network under study.

The mean number of users in the network is given by the following expression

$$N^{network} = \sum_{n=1}^{N} n \boldsymbol{\pi}_n \mathbf{e}.$$

The intensity of the output flow of successfully serviced users in the network can be found as

$$\lambda^{out} = \sum_{l=1}^{L} \sum_{n \in W_l/\{0\}} \boldsymbol{\pi}_n (I_V \otimes \mathcal{S}_n(\mathbf{p}^{(l)}))\mathbf{e} + \sum_{l=1}^{L-1} \sum_{n \in F_l} \sum_{i=0}^{1} \boldsymbol{\pi}(n, i)(I_V \otimes \mathcal{S}_n(\mathbf{p}^{(l+i)}))\mathbf{e}.$$

The output rate of users successfully serviced in the network from the k-th node is equal to

$$\lambda_k^{out} = \sum_{l=1}^{L} \sum_{n \in W_l/\{0\}} \boldsymbol{\pi}_n (I_V \otimes \mathcal{S}_n(\mathbf{p}^{(l,k)}))\mathbf{e} + \sum_{l=1}^{L-1} \sum_{n \in F_l} \sum_{i=0}^{1} \boldsymbol{\pi}(n, i)(I_V \otimes \mathcal{S}_n(\mathbf{p}^{(l+i,k)}))\mathbf{e}, \; k = \overline{1, K},$$

where $\mathbf{p}^{(l,k)}$ is a column vector of dimension K that has all zero entries except the k-th entry $(\mathbf{p}^{(l,k)})_k$, which is equal to $p_k^{(l)}$. The matrices $\mathcal{S}_n(\mathbf{p}^{(l,k)})$ can be found using the same algorithm as for the matrices $\mathcal{S}_n(\mathbf{p}^{(l)})$.

The average quantity of users in the k-th node, $k = \overline{1, K}$, is given by the following expression

$$N_k^{node} = \sum_{l=1}^{L} \sum_{n \in W_l/\{0\}} \boldsymbol{\pi}_n (I_V \otimes \mathcal{J}_n(\mathbf{b}^{(k)}))\mathbf{e} + \sum_{l=1}^{L-1} \sum_{n \in F_l} \sum_{i=0}^{1} \boldsymbol{\pi}(n, i)(I_V \otimes \mathcal{J}_n(\mathbf{b}^{(k)}))\mathbf{e}$$

where the matrices $\mathcal{J}_n(\mathbf{b})$, $n = \overline{1, N}$, for the vectors $\mathbf{b} = (b_1, \ldots, b_K)$, which take values from the set $\mathbf{b} \in \{\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(K)}\}$, can be found as

$$\mathcal{J}_n(\mathbf{b}) = \mathcal{J}_n^{(K-1)}(\mathbf{b}), \; n = \overline{1, N},$$

with the matrices $\mathcal{J}_n^{(K-1)}(\mathbf{b})$, $n = \overline{1, N}$, recursively obtained as

$$\mathcal{J}_n^{(0)}(\mathbf{b}) = nb_K,$$

$$\mathcal{J}_n^{(k)}(\mathbf{b}) = \begin{pmatrix} nb_{K-k}I & O & \cdots & O \\ \mathcal{J}_1^{(k-1)}(\mathbf{b}) & (n-1)b_{K-k}I & \cdots & O \\ O & \mathcal{J}_2^{(k-1)}(\mathbf{b}) & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & b_{K-k}I \\ O & O & \cdots & \mathcal{J}_n^{(k-1)}(\mathbf{b}) \end{pmatrix}, \; k = \overline{1, K-1}, \; n = \overline{1, N}.$$

The average quantity of busy servers in the k-th node, $k = \overline{1, K}$, is obtained using the formula

$$N_k^{serv} = \sum_{l=1}^{L} \sum_{n \in W_l/\{0\}} \boldsymbol{\pi}_n (I_V \otimes \mathcal{S}_n(\mathbf{b}^{(k)}))\mathbf{e} + \sum_{l=1}^{L-1} \sum_{n \in F_l} \sum_{i=0}^{1} \boldsymbol{\pi}(n, i)(I_V \otimes \mathcal{S}_n(\mathbf{b}^{(k)}))\mathbf{e}.$$

The average quantity of busy servers in the network is given by the following expression

$$N^{serv} = \sum_{k=1}^{K} N_k^{serv}.$$

The average quantity of users in the k-th node's buffer, $k = \overline{1, K}$, is given by the following expression

$$N_k^{buf} = \sum_{l=1}^{L} \sum_{n \in W_l/\{0\}} \boldsymbol{\pi}_n (I_V \otimes \mathcal{I}_n(\mathbf{b}^{(k)}))\mathbf{e} + \sum_{l=1}^{L-1} \sum_{n \in F_l} \sum_{i=0}^{1} \boldsymbol{\pi}(n, i)(I_V \otimes \mathcal{I}_n(\mathbf{b}^{(k)}))\mathbf{e} = N_{node}^{(k)} - N_{serv}^{(k)}.$$

The average quantity of users waiting in all buffers of the network can be calculated using the formula

$$N^{buf} = \sum_{k=1}^{K} N_k^{buf}.$$

The probability that the network operates in the l-th regime at an arbitrary epoch is determined by the following expression

$$P_l^{regime} = \sum_{n \in W_l} \boldsymbol{\pi}_n \mathbf{e} + \sum_{n \in F_l, \, l \neq L} \boldsymbol{\pi}(n, 0)\mathbf{e} + \sum_{n \in F_{l-1}, \, l \neq 1} \boldsymbol{\pi}(n, 1)\mathbf{e}, \; l = \overline{1, L}.$$

The intensity of the regime increase is given by the formula

$$\phi^+ = \sum_{l=1}^{L-1} \boldsymbol{\pi}(L_l^+, 0) \sum_{k=1}^{K} H_k \otimes I_{J_{L_l^+}}.$$

The intensity of the regime decrease is given by the following expression

$$\phi^- = \sum_{l=1}^{L-1} \Phi_l = \phi^+$$

where

$$\Phi_l = \begin{cases} \boldsymbol{\pi}(L_l^- + 1, 1)(I_V \otimes (\mathcal{I}_{L_l^-+1}(\boldsymbol{\beta}) + \mathcal{S}_{L_l^-+1}(\mathbf{p}^{(l+1)})) & \text{if } L_l^- \neq L_l^+, \\ \boldsymbol{\pi}_{L_l^-+1}(I_V \otimes (\mathcal{I}_{L_l^-+1}(\boldsymbol{\beta}) + \mathcal{S}_{L_l^-+1}(\mathbf{p}^{(l+1)})) & \text{if } L_l^- = L_l^+. \end{cases}$$

The average intensity of regime switching is equal to

$$\phi = \phi^+ + \phi^- = 2\phi^+ = 2\phi^-.$$

The probability of an arbitrary user loss upon arrival caused by the residence of N users in the network is calculated by the expression

$$P^{ent-loss} = \lambda^{-1}\boldsymbol{\pi}_N((H - H_0) \otimes I_{J_N})\mathbf{e}.$$

The probability of an arbitrary type-k user loss upon arrival due to the residence of N users in the network is computed as

$$P_k^{ent-loss} = \lambda_k^{-1}\boldsymbol{\pi}_N(H_k \otimes I_{J_N})\mathbf{e},\ k = \overline{1, K}.$$

The probability of an arbitrary user loss upon arrival at the k-th node due to the residence of N users in the network is given by the following expression

$$P_k^{ent-loss-arb} = \lambda^{-1}\boldsymbol{\pi}_N(H_k \otimes I_{J_N})\mathbf{e},\ k = \overline{1, K}.$$

The probability of an arbitrary user loss due to impatience is equal to

$$\begin{aligned} P^{imp-loss} &= \lambda^{-1}\sum_{k=1}^{K} N_k^{buf}\beta_k \\ &= \lambda^{-1}\Big(\sum_{l=1}^{L}\sum_{n\in W_l/\{0\}} \boldsymbol{\pi}_n(I_V \otimes \mathcal{I}_n(\boldsymbol{\beta}))\mathbf{e} + \sum_{l=1}^{L-1}\sum_{n\in F_l}\sum_{i=0}^{1}\boldsymbol{\pi}(n,i)(I_V \otimes \mathcal{I}_n(\boldsymbol{\beta}))\mathbf{e}\Big). \end{aligned}$$

The probability of an arbitrary user loss due to impatience in the k-th node is computed as

$$\begin{aligned} P_k^{imp-loss} &= \lambda^{-1}N_k^{buf}\beta_k \\ &= \lambda^{-1}\Big(\sum_{l=1}^{L}\sum_{n\in W_l/\{0\}} \boldsymbol{\pi}_n(I_V \otimes \mathcal{I}_n(\boldsymbol{\beta}^{(k)}))\mathbf{e} + \sum_{l=1}^{L-1}\sum_{n\in F_l}\sum_{i=0}^{1}\boldsymbol{\pi}(n,i)(I_V \otimes \mathcal{I}_n(\boldsymbol{\beta}^{(k)}))\mathbf{e}\Big),\ k = \overline{1, K}, \end{aligned}$$

where $\boldsymbol{\beta}^{(k)}$ is a column vector of size K with all zero entries except the k-th entry $(\boldsymbol{\beta}^{(k)})_k$, which is equal to β_k.

The probability of an arbitrary user loss is given by the following expression

$$P^{loss} = P^{ent-loss} + P^{imp-loss} = 1 - \frac{\lambda_{out}}{\lambda}.$$

Controlling the accuracy of the calculation of the stationary distribution of the network states is made easier by the existence of two distinct expressions for computing the probability P^{loss} and the average intensity of regime switching.

An arbitrary user's loss probability in the k-th node is determined as

$$P_k^{loss} = P_k^{ent-loss-arb} + P_k^{imp-loss},\ k = \overline{1, K}.$$

An arbitrary user's probability of receiving successful service within the network is determined by the formula

$$P^{succ} = 1 - P^{loss} = \frac{\lambda^{out}}{\lambda}.$$

5. Numerical example

Let us show the numerical example that verifies the viability of the suggested methods and formulas, and partially highlights the impact of variation of the thresholds on the value of the key performance indicators of the system and the potential to apply the outcomes to managerial objectives.

In this example, we examine a queueing network with $K = 3$ nodes.

The MMAP flow of users arriving at the network is determined by the following matrices:

$$H_0 = \begin{pmatrix} -9.3 & 0.3 \\ 0.3 & -2.7 \end{pmatrix}, H_1 = \begin{pmatrix} 3.3 & 0.03 \\ 0.009 & 0.579 \end{pmatrix},$$

$$H_2 = \begin{pmatrix} 2.4 & 0.15 \\ 0.012 & 1.2 \end{pmatrix}, H_3 = \begin{pmatrix} 3.06 & 0.06 \\ 0 & 0.6 \end{pmatrix}.$$

The average arrival rate for this arrival flow is $\lambda = 4.8606$. The average arrival rate λ_k and the coefficients of variation $c_{var}^{(k)}$ and correlation $c_{cor}^{(k)}$ of successive inter-arrival times to the k-th node, $k = 1, 2, 3$, have the following values:

$$\lambda_1 = 1.6103,\ c_{var}^{(1)} = 1.77393\ c_{cor}^{(1)} = 0.181652,$$

$$\lambda_2 = 1.7108,\ c_{var}^{(2)} = 2.05727\ c_{cor}^{(2)} = 0.148899,$$

$$\lambda_3 = 1.5395,\ c_{var}^{(3)} = 1.16264\ c_{cor}^{(3)} = 0.0462668.$$

We assume that there are $L = 3$ possible service regimes of the network operation. Under the first regime, the service times in the nodes are exponentially distributed with parameters

$$\mu_1^{(1)} = 1.5, \mu_2^{(1)} = 1, \mu_3^{(1)} = 0.9,$$

respectively. Under the second regime, the parameters of the exponential distribution of the service time are

$$\mu_1^{(2)} = 2\mu_1^{(1)}, \mu_2^{(2)} = 2\mu_2^{(1)}, \mu_3^{(2)} = 2\mu_3^{(1)},$$

and under the third regime, the parameters are

$$\mu_1^{(3)} = 3\mu_1^{(1)}, \mu_2^{(3)} = 3\mu_2^{(1)}, \mu_3^{(3)} = 3\mu_3^{(1)}.$$

The transition probabilities of users after the service completion in the nodes are defined as

$$p_{1,0} = 3/5,\ p_{1,2} = 2/15,\ p_{1,3} = 4/15,\ p_{2,0} = 0.7,$$

$$p_{2,1} = 0.1,\ p_{2,3} = 0.2,\ p_{3,0} = 2/3,\ p_{3,1} = 2/9,\ p_{3,2} = 1/9.$$

The rates of the users' departure from the buffers of the nodes due to impatience are defined as follows:

$$\beta_1 = 0.01, \beta_2 = 0.02, \beta_3 = 0.015.$$

In this numerical example, we assume that up to $N = 40$ users can obtain service in the network at the same time. The regime switching is defined by the four parameters L_1^-, L_2^-, L_1^+ and L_2^+.

The purpose of the numerical example is to show the dependence of the main network's performance measures on the switching parameters. However, there is no opportunity to build 5 D figures. Therefore, for better visualization of the results, let us fix the thresholds defining the rule of the switching between the first and the second regimes as $L_1^- = 5$ and $L_1^+ = 10$ and vary the thresholds L_2^- and L_2^+, defining the rule of the switching between the second and third regimes as follows: The threshold L_2^+ varies in the interval $[L_1^+ + 1, N)$, and the threshold L_2^- varies over the interval $[L_1^+ + 1, L_2^+]$ with the same step 1.

Figures 2 and 3 illustrate the dependence of the average quantity $N^{network}$ and the average total number N^{buf} of users in buffers on the parameters L_2^- and L_2^+.

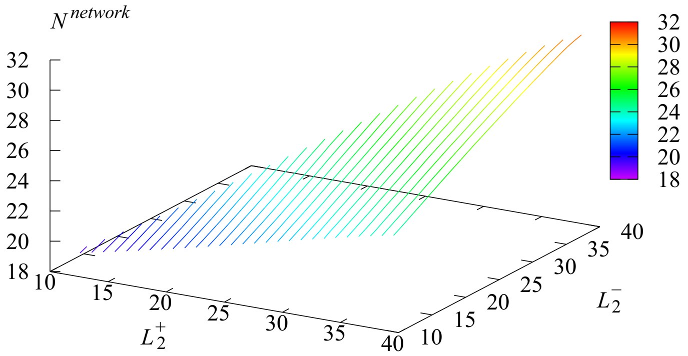

Figure 2. $N^{network}$ as function of L_2^- and L_2^+.

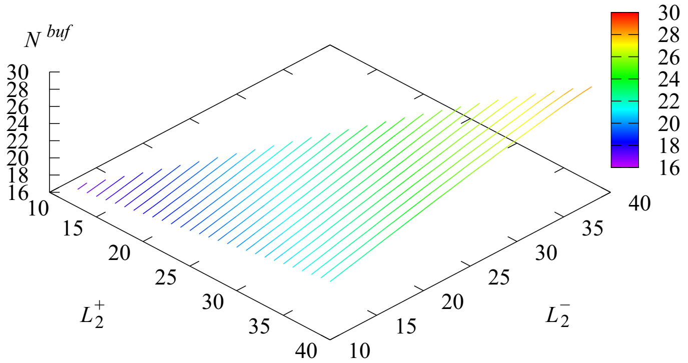

Figure 3. N^{buf} as function of L_2^- and L_2^+.

Complementary to Figure 2, the dynamics of $N^{network}$ are illustrated in Table 1 where values of $N^{network}$ are presented for values of L_2^- and L_2^+ in some smaller range.

It is seen from these figures that the minimal values of $N^{network}$ and N^{buf} are achieved for small values of the thresholds L_2^- and L_2^+. When these thresholds increase (this means that the third regime is used only for a larger number of users in the network), the values of $N^{network}$ and N^{buf} increase quite sharply.

This is easily understandable, as the service rate during the use of the third regime is three times higher than during the use of the first regime and is 1.5 times higher than during the use of the second regime.

Table 1. Values of $N^{network}$ for different values of L_2^- and L_2^+.

L_2^+ \ L_2^-	11	12	13	14	15	16	17	18	19	20
11	19.089									
12	19.256	19.451								
13	19.422	19.627	19.834							
14	19.587	19.801	20.019	20.237						
15	19.753	19.973	20.199	20.428	20.655					
16	19.921	20.146	20.378	20.615	20.853	21.088				
17	20.090	20.319	20.556	20.799	21.044	21.290	21.532			
18	20.263	20.494	20.735	20.981	21.233	21.486	21.738	21.986		
19	20.438	20.673	20.915	21.164	21.419	21.678	21.938	22.195	22.447	
20	20.617	20.853	21.097	21.349	21.606	21.868	22.132	22.397	22.659	22.914
21	20.798	21.037	21.283	21.535	21.794	22.057	22.325	22.594	22.863	23.128
22	20.982	21.222	21.470	21.724	21.983	22.248	22.516	22.788	23.061	23.333
23	21.168	21.410	21.659	21.914	22.174	22.439	22.708	22.981	23.256	23.532
24	21.355	21.599	21.850	22.106	22.367	22.632	22.902	23.174	23.450	23.727
25	21.543	21.790	22.042	22.299	22.561	22.827	23.096	23.368	23.643	23.921
26	21.733	21.981	22.235	22.493	22.756	23.022	23.291	23.563	23.838	24.114
27	21.923	22.172	22.428	22.688	22.951	23.218	23.487	23.759	24.033	24.308
28	22.112	22.364	22.621	22.882	23.146	23.414	23.683	23.955	24.228	24.502
29	22.302	22.555	22.813	23.075	23.3413	23.609	23.878	24.150	24.422	24.696
30	22.490	22.745	23.005	23.268	23.534	23.802	24.073	24.344	24.616	24.889
31	22.677	22.934	23.194	23.459	23.726	23.995	24.265	24.536	24.808	25.08
32	22.863	23.120	23.382	23.647	23.915	24.184	24.455	24.726	24.997	25.269
33	23.046	23.304	23.567	23.833	24.101	24.371	24.642	24.913	25.184	25.455
34	23.226	23.485	23.749	24.016	24.284	24.554	24.825	25.096	25.367	25.637
35	23.402	23.663	23.927	24.194	24.463	24.733	25.004	25.275	25.545	25.814
36	23.574	23.835	24.101	24.368	24.637	24.907	25.178	25.448	25.718	25.986
37	23.741	24.003	24.268	24.535	24.805	25.075	25.345	25.615	25.884	26.152
38	23.901	24.163	24.429	24.697	24.966	25.235	25.505	25.774	26.042	26.309
39	24.054	24.317	24.582	24.85	25.118	25.387	25.656	25.925	26.192	26.457

Figures 4–6 illustrate the dependence of the probabilities P_l^{regime} on the fact that, at any arbitrary moment, the network operates in the l-th regime, $l = 1, 2, 3$, on the parameters L_2^- and L_2^+.

The maximum value of P_1^{regime} is achieved for small values of L_2^- and L_2^+ because such small values imply a more frequent use of the fastest, the third, regime of operation (this is confirmed by Figure 6) and higher chances that the number of users in the network will drop below the value $L_1^- + 1$ and, therefore, the first regime will be used. The maximum value of P_2^{regime} is achieved for large values of L_2^- and L_2^+ because the second regime is used until the number of users in the network drops below the value $L_2^- + 1$, which implies the longer use of the second regime.

Figure 7 illustrates the dependence of the average intensity ϕ of regime switching on the parameters

L_2^- and L_2^+.

Figures 8–11 illustrate the dependence of the probabilities $P^{ent-loss}$ of a user loss at the entrance to the network and the probabilities $P_l^{ent-loss-arb}$ of a user loss at the entrance to the l-th node of the network, $l = 1, 2, 3$, on the parameters L_2^- and L_2^+.

Figures 12–15 illustrate the dependence of the probabilities $P^{imp-loss}$ of a user loss due to impatience in the network and the probabilities $P_l^{imp-loss}$ of a user loss due to impatience from the buffer of the l-th node of the network, $l = 1, 2, 3$, on the parameters L_2^- and L_2^+.

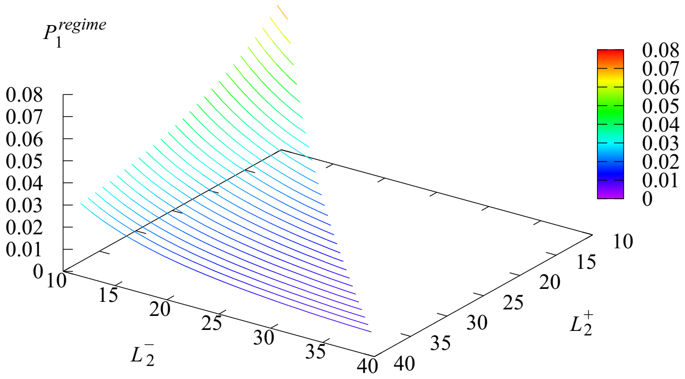

Figure 4. P_1^{regime} as function of L_2^- and L_2^+

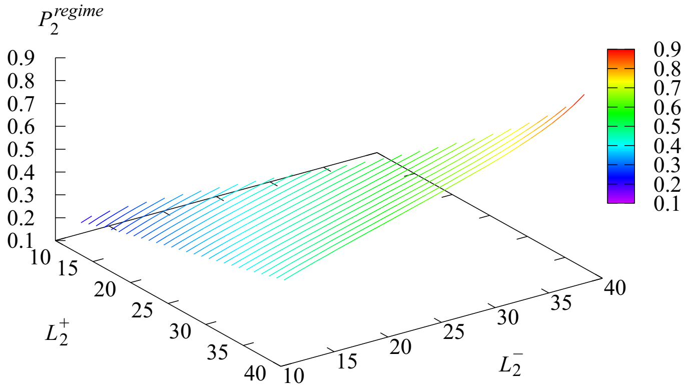

Figure 5. P_2^{regime} as function of L_2^- and L_2^+

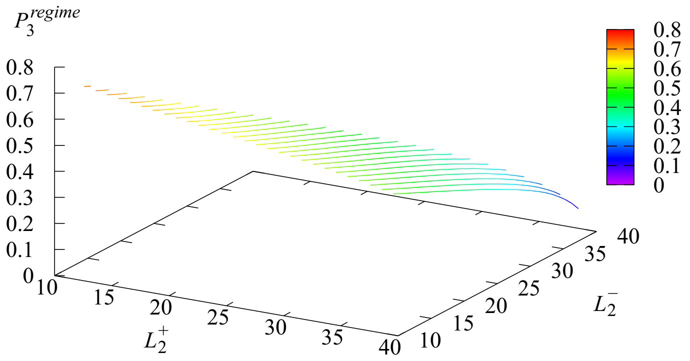

Figure 6. P_3^{regime} as function of L_2^- and L_2^+

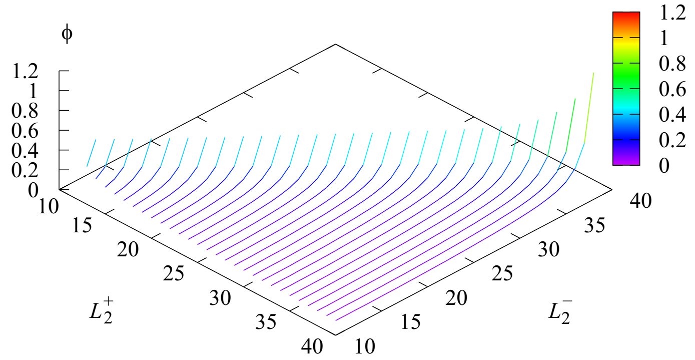

Figure 7. ϕ as function of L_2^- and L_2^+

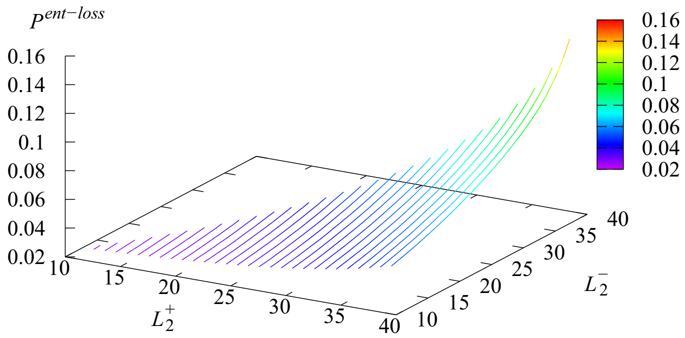

Figure 8. $P^{ent-loss}$ as function of L_2^- and L_2^+

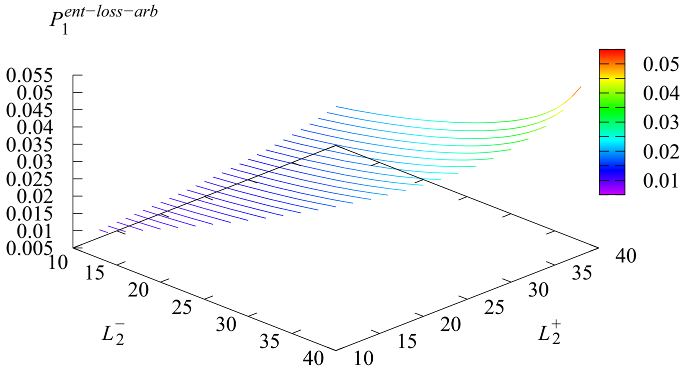

Figure 9. $P_1^{ent-loss-arb}$ as function of L_2^- and L_2^+

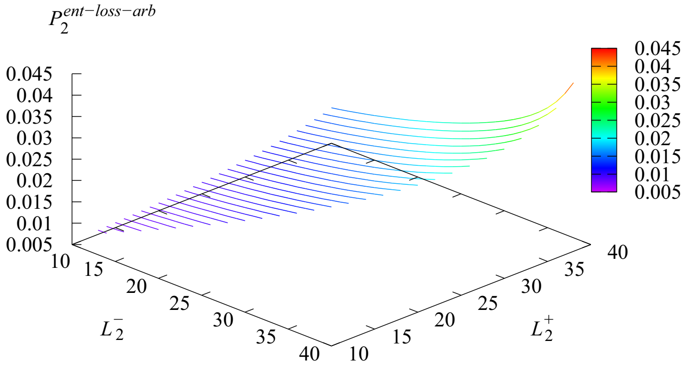

Figure 10. $P_2^{ent-loss-arb}$ as function of L_2^- and L_2^+

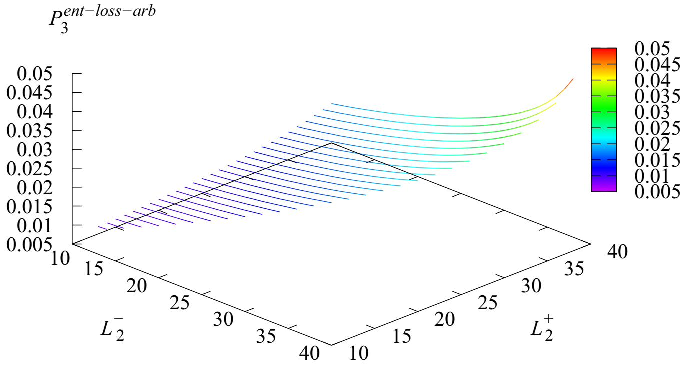

Figure 11. $P_3^{ent-loss-arb}$ as function of L_2^- and L_2^+

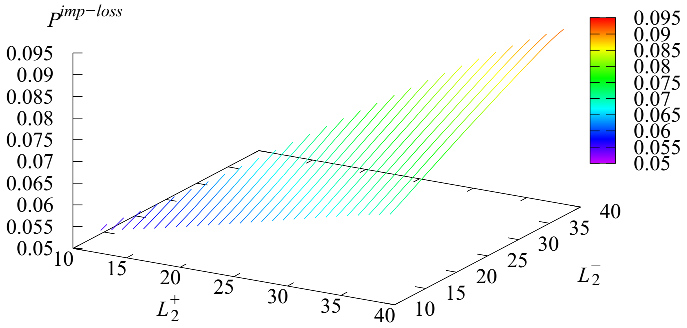

Figure 12. $P^{imp-loss}$ as function of L_2^- and L_2^+

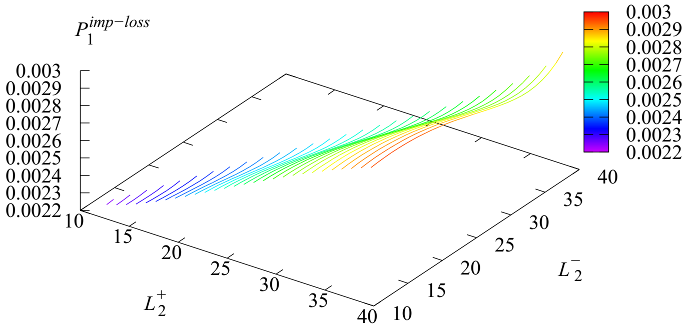

Figure 13. $P_1^{imp-loss}$ as function of L_2^- and L_2^+

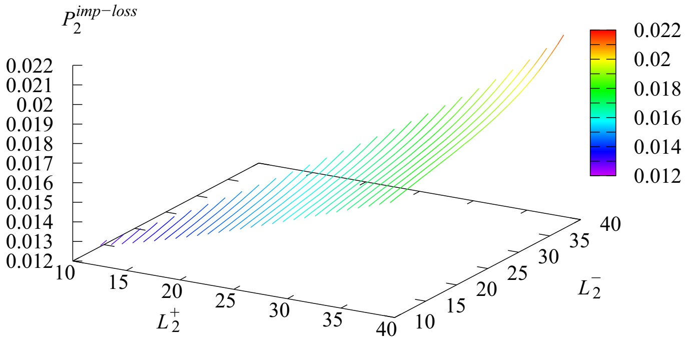

Figure 14. $P_2^{imp-loss}$ as function of L_2^- and L_2^+

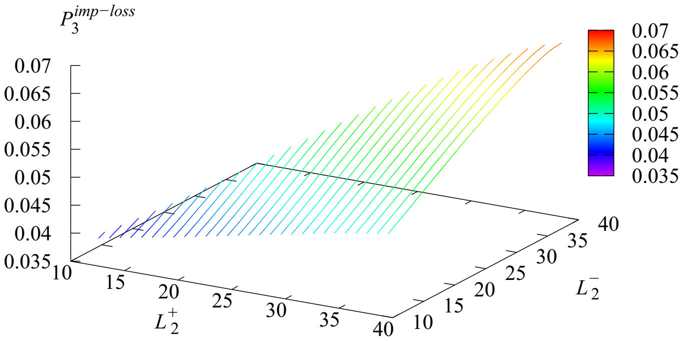

Figure 15. $P_3^{imp-loss}$ as function of L_2^- and L_2^+

Figure 16 illustrates the dependence of the loss probability P^{loss} of an arbitrary user (due to all reasons) on the parameters L_2^- and L_2^+.

Complementary to Figure 16, the dynamics of P^{loss} are illustrated in Table 2.

Because we have fixed $L_1^+ = 10$, and the thresholds L_2^- and L_2^+ have to satisfy the inequalities $L_1^+ < L_2^- \leq L_2^+ < N$, the possible values of the thresholds L_2^- and L_2^+ range from 11 to 39. The loss probability P^{loss} reaches its minimum value of 0.07887 when $L_2^- = L_2^+ = 11$. When $L_2^- = L_2^+ = 39$, the maximum value of the loss probability P^{loss} is reached and equals 0.23454. This fact is obvious because when $L_2^- = L_2^+ = 11$, the network starts operation in the fastest, the third, service regime as soon as possible. When $L_2^- = L_2^+ = 39$, the third regime is switched on only when the number of users in the system is equal to the maximum admissible value of 40. Therefore, many users are lost at the entrance to the network and due to impatience. However, when deciding on the selection of the values of the thresholds, it is necessary to take into account that the use of faster service regimes by default is more costly compared to slower service regimes. Also, it is undesirable to frequently change service regimes because such a change in a real-world network can require some expenditures related to switching to another equipment or inviting or dismissing staff.

Therefore, optimization of the operation of the network requires an exact definition of the cost criterion. Let us assume that the following cost criterion is used to define the network's operational quality:

$$E = E(L_2^-, L_2^+) = a\lambda^{out} - b\lambda P^{ent-loss} - c\lambda P^{imp-loss} - \sum_{l=1}^{L} e_l P_l^{regime} - d\phi. \tag{3}$$

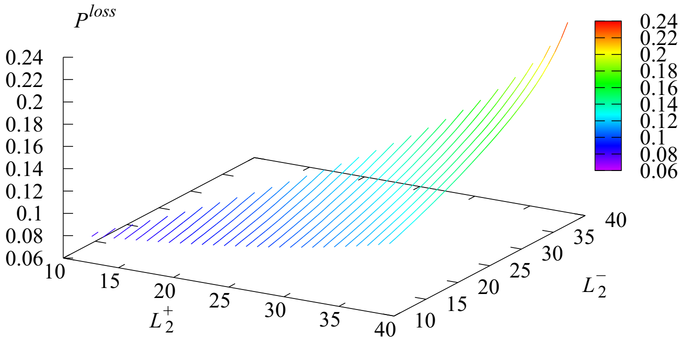

Figure 16. P^{loss} as function of L_2^- and L_2^+

Table 2. Values of P^{loss} for different values of L_2^- and L_2^+

L_2^- L_2^+	11	12	13	14	15	16	17	18	19	20
11	0.0788									
12	0.0797	0.0807								
13	0.0806	0.0817	0.0828							
14	0.0815	0.0826	0.0838	0.0850						
15	0.0824	0.0836	0.0848	0.0861	0.0873					
16	0.0833	0.0845	0.0858	0.0871	0.0885	0.0899				
17	0.0843	0.0855	0.0869	0.0882	0.0896	0.0911	0.0925			
18	0.0853	0.0866	0.0879	0.0893	0.0908	0.0923	0.0938	0.0953		
19	0.0863	0.0876	0.0890	0.0905	0.0920	0.0935	0.0951	0.0967	0.0983	
20	0.0874	0.0888	0.0902	0.0916	0.0932	0.0947	0.0964	0.0981	0.0998	0.1015
21	0.0886	0.0899	0.0913	0.0928	0.0944	0.0960	0.0977	0.0994	0.1012	0.1030
22	0.0897	0.0911	0.0926	0.0941	0.0957	0.0973	0.0990	0.1008	0.1026	0.1045
23	0.0909	0.0924	0.0939	0.0954	0.0970	0.0987	0.1004	0.1022	0.1041	0.1060
24	0.0922	0.093	0.0952	0.0968	0.0984	0.1001	0.1019	0.1037	0.1056	0.1075
25	0.0935	0.0950	0.0966	0.0982	0.0998	0.1016	0.1034	0.1052	0.1071	0.1091
26	0.0949	0.0964	0.0980	0.0996	0.1013	0.1031	0.1049	0.1068	0.1088	0.1108
27	0.0963	0.0979	0.0995	0.1012	0.1029	0.1047	0.1066	0.1085	0.1105	0.1125
28	0.0978	0.0994	0.1011	0.1028	0.1045	0.1064	0.1083	0.1102	0.1122	0.1143
29	0.0994	0.1010	0.1027	0.1044	0.1062	0.1081	0.1100	0.1120	0.1141	0.1162
30	0.1010	0.1027	0.1044	0.1062	0.1080	0.1099	0.1119	0.1139	0.1160	0.1182
31	0.1027	0.1044	0.1062	0.1080	0.1099	0.1118	0.1139	0.1159	0.1181	0.1203
32	0.1045	0.1062	0.1080	0.1099	0.1118	0.1138	0.1159	0.1180	0.1202	0.1225
33	0.1063	0.1081	0.1100	0.1119	0.1139	0.1159	0.1180	0.1202	0.1225	0.1248
34	0.1083	0.1101	0.1120	0.1140	0.1160	0.1181	0.1203	0.1225	0.1248	0.1272
35	0.1104	0.1122	0.1142	0.1162	0.1183	0.1205	0.1227	0.1250	0.1274	0.1298
36	0.1125	0.1145	0.1165	0.1186	0.1207	0.1229	0.1252	0.1276	0.1300	0.1325
37	0.1148	0.1168	0.1189	0.1211	0.1233	0.1256	0.1279	0.1303	0.1328	0.1354
38	0.1173	0.1194	0.1215	0.1237	0.1260	0.1283	0.1308	0.1333	0.1359	0.1385
39	0.1199	0.1220	0.1242	0.1265	0.1289	0.1313	0.1338	0.1364	0.1391	0.1418

Here, a is a profit obtained by the system via the service of one user, b is a penalty paid by the network for one user loss upon arrival, c is a penalty paid by the network for one user loss due to impatience, e_l, $l = \overline{1, L}$, is the cost of maintaining the l-th operation regime per unit time, and d is a charge paid by the network for one switch of an operation regime.

Thus, the cost criterion E represents the average network's revenue per unit of time. Our aim is to find the values of the thresholds L_2^- and L_2^+ providing the maximum to the function $E(L_2^-, L_2^+)$.

The cost coefficients in this numerical example are fixed at the following values:

$$a = 3,\ b = 3,\ c = 6,\ e_1 = 1,\ e_2 = 2,\ e_3 = 8,\ d = 0.5.$$

Figure 17 shows how the thresholds L_2^- and L_2^+ affect the cost criteria E.

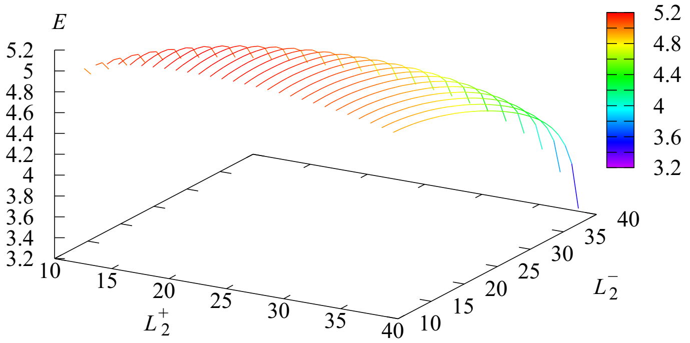

Figure 17. E as function of L_2^- and L_2^+

When $L_2^+ = 20$ and $L_2^- = 15$, the cost criterion reaches its optimal value of $E^* = 5.19909$. Thus, to obtain the maximal revenue under the fixed above values of the network parameters, it is necessary to switch from the second to the third service regime when the number of users in the network becomes equal to $(L_2^+) + 1 = 21$ and switch back to the second service regime when the number of users drops to $L_2^- = 15$.

All figures presented above illustrate the dependence of the cost criterion E on the thresholds L_2^- and L_2^+ under the fixed values of the thresholds $L_1^- = 5$ and $L_1^+ = 10$. Let us now assume that $L_1^- = L_1^+ = L_1$ and $L_2^- = L_2^+ = L_2$, i.e., the hysteresis strategy turns into the threshold strategy. This means that if the number n_t of users in the network does not exceed L_1, then the network operates in the first regime. If the number n_t belongs to the interval $(L_1 + 1, L_2]$, then the network operates in the second regime. If the number n_t exceeds L_2, then the network operates in the third regime.

Figures 18 and 19 illustrate the dependence of the loss probabilities $P^{ent-loss}$ of an arbitrary user upon arrival and $P^{imp-loss}$ of an arbitrary user due to impatience on the parameters L_1 and L_2.

Figure 20 illustrates the dependence of the cost criterion $E(L_1, L_2)$ on the parameters L_1 and L_2.

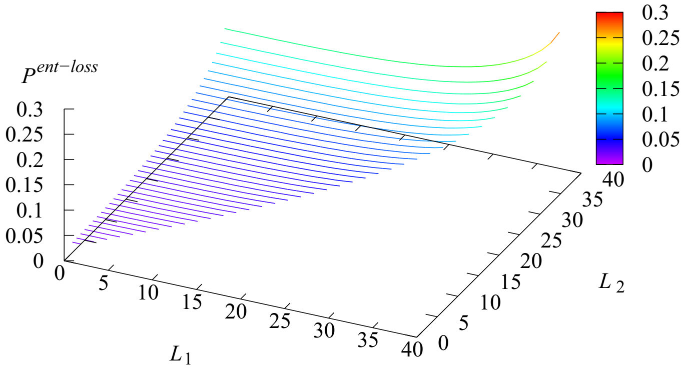

Figure 18. $P^{ent-loss}$ as function of L_1 and L_2

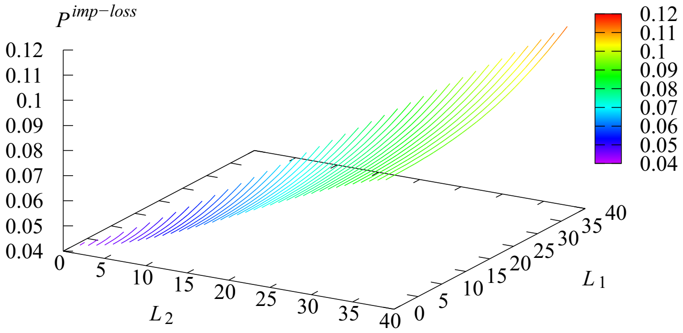

Figure 19. $P^{imp-loss}$ as function of L_1 and L_2

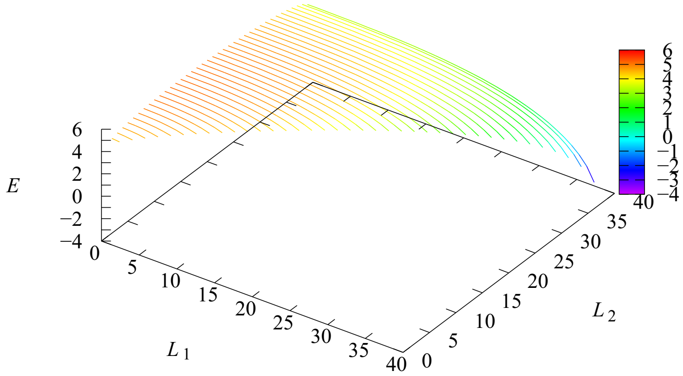

Figure 20. The cost criterion $E(L_1, L_2)$ as function of L_1 and L_2

The optimal values of the thresholds are as follows: $L_1 = 0$ and $L_2 = 15$. This means that the network uses the first regime only when it is empty; once a user arrives, it is serviced in the second regime. When the number of users in the network reaches the value of 16, the network starts operation in the third regime and maintains this regime until the number of users in the network drops to the value of 15. The maximal value of the cost criterion is 5.13969.

Let us return to the hysteresis-type control by the network operation. A numerical solution to the problem of determining the optimal value of the cost criterion under the network's fixed parameters can be found using the results obtained for the computation of the stationary distribution of the network states, the primary performance characteristics, and the cost criterion value. In our example, the value of the cost criterion is the function of four thresholds, L_1^-, L_1^+, L_2^-, and L_2^+. It can be computed that the maximum value of the cost criterion is equal to 5.31252 and is achieved for the following values of the thresholds: $L_1^- = 0, L_1^+ = 2, L_2^- = 13, L_2^+ = 18$. The achieved value of 5.31252 of the cost criterion is higher than the optimal value of 5.13969 of the criterion under the use of the optimal threshold strategy due to the more seldom switching of regimes and the presence of a charge for the switch of the regimes in the cost criterion.

In a more general case, when the number L of available regimes is higher than two (and the number of the thresholds is $2L$), the problem of finding the maximal value of the criterion and the optimal values of the corresponding thresholds can be deeply complicated by the existence of a huge number of possible combinations of the threshold values. Therefore, this problem deserves a separate consideration. The use of some derivative-free methods of optimization, see, e.g., [45, 46], can be recommended. As mentioned above, the value of our results consists in providing the possibility to exactly compute the value of the cost criterion for any fixed set of control strategies during the implementation of the search for the optimal values of the thresholds.

It is worth noting that we assumed that the values of the service rates $\mu_k^{(l)},\ k = \overline{1, K},\ l = \overline{1, L}$, in the nodes of the network under the fixed regimes of the network operation are fixed. In potential real-world applications, these values can also not be fixed but have to be chosen from some set. Our results can be used to optimize the selection of these rates' values and the corresponding threshold values.

Remark 3. *Described above computations were implemented using Wolfram Mathematica on a Lenovo notebook with an Intel(R) Core(TM) i7-1165G7 2.80GHz and 16 GB RAM. Running time for computation of the optimal value of the cost criterion $E(L_2^-, L_2^+)$ defined by formula (3) was equal to 8419 seconds, i.e., 18 seconds for one point (L_2^-, L_2^+) (the total number of points is 465). Because this computation time was acceptable for preparation of the presented examples, no optimization of the code was made. Computation time can be significantly reduced via such an optimization and the use of a more powerful notebook or PC.*

A quite long computation time is explained by the large size of the generator Q. As mentioned above in the simplest case of the threshold strategy, this size is equal to $V \sum\limits_{n=0}^{N} J_n$, where the numbers $J_n,\ n = \overline{0, N}$, are defined by formula (2). In the considered example of $SOQN$, where we assume $K = 3$ nodes and admission of up to $N = 40$ customers to the network simultaneously, the number J_N is equal to 861. If we consider the network consisting of four nodes and decreased N to 15, we will have about the same (816) size of the block J_N and a similar computation time.

6. Conclusions

We considered a user loss SOQN with the bursty MMAP-type arrival process, single-server nodes with a controlled service regime, a hysteresis-type control policy, and impatient users. Under the fixed parameters of the control policy, the stationary behavior of this SOQN is described by a multidimensional Markov chain, whose components define the total number of users in the network, the used service regime (if it is not uniquely defined by the number of users in the network), the state of the underlying process of the MMAP, and the number of users in each node of the network. The generator of this chain is obtained as a block tridiagonal matrix. Formulas for computation of the key performance measures of the network are derived. Numerical illustrations of the algorithmic and analytical results obtained are provided.

The results can be applied to managerial objectives, such as the selection of possible variants of service organization in the nodes, including the choice of the suitable equipment and the corresponding staff; routing of users in the network; pricing; and the optimal dynamical scheduling of the variants of service organization depending on the current load of the network.

The results can be generalized into several directions, such as the consideration of back-ordering SOQNs having an infinite or finite buffer for storing users who did not succeed in entering the core network upon arrival because it was completely busy; an account of possible user's impatience during waiting in this buffer; the presence of an orbit for user retrials; or the arrival and impatience rate control. The results from [47] are planned to be used for implementing these generalizations.

Author contributions

Ciro D'Apice: Conceptualization, Methodology, Validation, Investigation, Writing-original draft, Writing-review & editing, Project administration; Alexander Dudin: Conceptualization, Methodology, Formal analysis, Investigation, Writing-original draft, Writing-review & editing, Project administration; Sergei Dudin: Methodology, Software, Formal analysis, Investigation, Supervision, Writing-original draft; Rosanna Manzo: Conceptualization, Software, Validation, Formal analysis, Investigation, Writing-original draft, Writing-review & editing. All authors have read and approved the final version of the manuscript for publication.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

Rosanna Manzo is the Guest Editor of special issue "Managing complex systems by simulation and optimization techniques" for AIMS Mathematics. Rosanna Manzo was not involved in the editorial review and the decision to publish this article.

References

1. D. Roy, Semi-open queuing networks: a review of stochastic models, solution methods and new research areas, *Int. J. Product. Res.*, **54** (2016), 1735–1752. https://doi.org/10.1080/00207543.2015.1056316
2. Y. Dallery, Approximate analysis of general open queuing networks with restricted capacity, *Perform. Evaluation*, **11** (1990), 209–222, https://doi.org/10.1016/0166-5316(90)90013-9
3. S. Otten, R. Krenzler, L. Xie, H. Daduna, K. Kruse, Analysis of semi-open queuing networks using lost customers approximation with an application to robotic mobile fulfilment systems, *OR Spectrum*, **44** (2022), 603–648. https://doi.org/10.1007/s00291-021-00662-9
4. B. Avi-Itzhak, D. P. Heyman, Approximate queuing models for multiprogramming computer systems, *Oper. Res.*, **21** (1973), 1212–1230. https://doi.org/10.1287/opre.21.6.1212
5. J. Jia, S. Heragu, Solving semi-open queuing networks, *Oper. Res.*, **57** (2009), 391–401. https://doi.org/10.1287/opre.1080.0627
6. B. Ekren, S. Heragu, A. Krishnamurthy, C. Malmborg, Matrix-geometric solution for semi-open queuing network model of autonomous vehicle storage and retrieval system, *Comput. Ind. Eng.*, **68** (2014), 78–86. https://doi.org/10.1016/j.cie.2013.12.002
7. J. Kim, A. Dudin, S. Dudin, C. Kim, Analysis of a semi-open queueing network with Markovian arrival process, *Perform. Evaluation*, **120** (2018), 1–19. https://doi.org/10.1016/j.peva.2017.12.005
8. C. Kim, S. Dudin, A. Dudin, K. Samouylov, Analysis of a semi-open queuing network with a state dependent marked Markovian arrival process, customers retrials and impatience, *Mathematics*, **7** (2019), 715. https://doi.org/10.3390/math7080715

9. C. Kim, S. Dudin, Analysis of semi-open queueing network with customer retrials, *J. Korean Inst. Indust. Eng.*, **45** (2019), 193–202. https://doi.org/10.7232/JKIIE.2019.45.3.193

10. S. Dudin, A. Dudin, R. Manzo, L. Rarità, Analysis of semi-open queueing network with correlated arrival process and multi-server nodes, *Oper. Res. Forum*, **5** (2024), 99. https://doi.org/10.1007/s43069-024-00383-z

11. M. Amjath, L. Kerbache, A. Elomri, J. M. Smith, Queueing network models for the analysis and optimisation of material handling systems: a systematic literature review, *Flex. Serv. Manuf. J.*, **36** (2024), 668–709. https://doi.org/10.1007/s10696-023-09505-x

12. J. Mao, J. Cheng, X. Li, H. Zhao, C. Lin, Modelling analysis of a four-way shuttle-based storage and retrieval system on the basis of operation strategy, *Appl. Sci.*, **13** (2023), 3306. https://doi.org/10.3390/app13053306

13. P. Legato, R. M. Mazza, Queueing networks for supporting container storage and retrieval, *Maritime Bus. Rev.*, **8** (2023) 301–317. https://doi.org/10.1108/MABR-01-2023-0009

14. H. Xie, T. J. Chaussalet, M. Rees, A semi-open queueing network approach to the analysis of patient flow in healthcare systems, In: *Proceedings of the 20th IEEE International Symposium on Computer-Based Medical Systems. IEEE CBMS 2007, Maribor, Slovenia, 20–22 June 2007 Los Alamitos, USA IEEE*, 2007, 719–724. https://doi.org/10.1109/CBMS.2007.12

15. P. Yang, G. Jin, G. Duan, Modelling and analysis for multi-deep compact robotic mobile fulfilment system, *Int. J. Product. Res.*, **60** (2022), 4727–4742. https://doi.org/10.1080/00207543.2021.1936264

16. L. Luo, N. Zhao, Y. Zhu, Y. Sun, A guiding DQN algorithm for automated guided vehicle pathfinding problem of robotic mobile fulfillment systems, *Comput. Indust. Eng.*, **178** (2023), 109112. https://doi.org/10.1016/j.cie.2023.109112

17. J. Lu, C. Ren, Y. Shao, J. Zhu, X. Lu, An automated guided vehicle conflict-free scheduling approach considering assignment rules in a robotic mobile fulfillment system, *Comput. Indust. Eng.*, **176** (2023), 108932. https://doi.org/10.1016/j.cie.2022.108932

18. G. Jiao, H. Li, M. Huang, Online joint optimization of pick order assignment and pick pod selection in robotic mobile fulfillment systems, *Comput. Indust. Eng.*, **175** (2023), 108856. https://doi.org/10.1016/j.cie.2022.108856

19. G. Tadumadze, J. Wenzel, S. Emde, F. Weidinger, R. Elbert, Assigning orders and pods to picking stations in a multi-level robotic mobile fulfillment system, *Flex. Serv. Manuf. J.*, **35** (2023), 1038–1075. https://doi.org/10.1007/s10696-023-09491-0

20. H. Li, H. Zhu, D. Xu, X. Lin, G. Jiao, Y. Song, et al., Dynamic task allocation based on auction in robotic mobile fulfilment system, *J. Indust. Manag. Optim.*, **19** (2023), 7600–7615. https://doi.org/10.3934/jimo.2023010

21. T. Lamballais, M. Merschformann, D. Roy, M. B. M. de Koster, K. Azadeh, L. Suhl, Dynamic policies for resource reallocation in a robotic mobile fulfillment system with time-varying demand, *European J. Oper. Res.*, **300** (2022), 937–952. https://doi.org/10.1016/j.ejor.2021.09.001

22. W. Chen, P. Wu, Y. Gong, Z. Zhang, K. Wang, The role of energy consumption in robotic mobile fulfillment systems: Performance evaluation and operating policies with dynamic priority, *Omega*, **130** (2025), 103168. https://doi.org/10.1016/j.omega.2024.103168

23. Y. Luo, M. Chen, L. Zeng, C. Zhang, Production system configuration design for an unmanned manufacturing factory, *Indust. Eng. Appl.*, **35** (2023), 13–22. https://doi.org/10.3233/ATDE230026

24. X. R. Chen, X. P. Liu, A. L. Yu, An integrated queuing network model for optimizing multi-level AVS/RS performance in multi-floor manufacturing environments, *IEEE Access*, **12** (2024), 181741–181755. https://doi.org/10.1109/ACCESS.2024.3507283

25. M. F. Neuts, A versatile Markovian point process, *J. Appl. Prob.*, **16** (1979), 764–779. https://doi.org/10.2307/3213143

26. D. Lucantoni, New results on the single server queue with a batch Markovian arrival process, *Commun. Stat. Stochast. Models*, **7** (1991), 1–46. https://doi.org/10.1080/15326349108807174

27. S. R. Chakravarthy, The batch Markovian arrival process: A review and future work, *Adv. Probab. Theory Stoch. Process*, **1** (2001), 21-49.

28. S. R. Chakravarthy, Introduction to matrix-analytic methods in queues 1: analytical and simulation approach-basics, In: *ISTE Ltd, London and John Wiley and Sons, New York*, 2022. https://doi.org/10.1002/9781394165421

29. S. R. Chakravarthy, Introduction to matrix-analytic methods in queues 2: analytical and simulation approach-queues and simulation, In: *ISTE Ltd, London and John Wiley and Sons, New York,* 2022. https://doi.org/10.1002/9781394174201

30. A. N. Dudin, V. I. Klimenok, V. M. Vishnevsky, *The theory of queuing systems with correlated flows*, Berlin: Springer, 2020. https://doi.org/10.1007/978-3-030-32072-0

31. M. Gonzalez, R. E. Lillo, J. Ramirez Cobo, Call center data modeling: a queueing science approach based on Markovian arrival processes, *Qual. Technol. Quant. Manag.*, 2024. http://doi.org/10.1080/16843703.2024.2371715

32. Q. M. He, Queues with marked customers, *Adv. Appl. Prob.*, **28** (1996), 567–587.

33. T. B. Crabill, Optimal control of a service facility with variable exponential service times and constant arrival rate, *Manag. Sci.*, **18** (1972), 560–566. https://doi.org/10.1287/mnsc.18.9.560

34. H. C. Tijms, On the optimality of a switch-over policy for controlling the queue size in a $M/G/1$ queue with variable service rate, *Lecture Notes Comput. Sci.*, **40** (1976), 736–742. https://doi.org/10.1007/3-540-07622-0_506

35. A. Dudin, Optimal multithreshold control for a $BMAP/G/1$ queue with N service modes, *Queueing Syst.*, **30** (1998), 273–287. https://doi.org/10.1023/A:1019121222439

36. C. S. Kim, V. Klimenok, A. Birukov, A. Dudin, Optimal multi-threshold control by the $BMAP/SM/1$ retrial system, *Ann. Oper. Res.*, **141** (2006), 193–210. https://doi.org/10.1007/s10479-006-5299-3

37. R. D. Nobel, H. C. Tijms, Optimal control for an $M^X/G/1$ queue with two service modes, *European J. Oper. Res.*, **113** (1999), 610–619. https://doi.org/10.1016/S0377-2217(98)00085-X

38. A. Dudin, Optimal control for an $M^X/G/1$ queue with two operation modes, *Prob. Eng. Inform. Sci.*, **11** (1997), 255–265. 10.1017/S0269964800004794

39. A. N. Dudin, S. Nishimura, Optimal control for a $BMAP/G/1$ queue with two service modes, *Math. Prob. Eng.*, **5** (1999), 255–273. https://doi.org/10.1155/S1024123X99001088

40. V. Ramaswami, D. M. Lucantoni, Algorithms for the multi-server queue with phase type service, *Stochastic Models*, **1** (1985), 393–417. https://doi.org/10.1080/15326348508807020

41. S. Sharma, R. Kumar, B. S. Soodan, P. Singh, Queuing models with customers' impatience: a survey, *Int. J. Math. Oper. Res.*, **26** (2023), 523–547. https://doi.org/10.1504/IJMOR.2023.135546

42. A. Graham, *Kronecker products and matrix calculus with applications*, New York: Courier Dover Publications, 2018.

43. C. S. Kim, S. A. Dudin, O. S. Taramin, J. Baek, Queueing system $MAP/PH/N/N+R$ with impatient heterogeneous customers as a model of call center, *Appl. Math. Model.*, **37** (2013), 958–976. https://doi.org/10.1016/j.apm.2012.03.021

44. H. Baumann, W. Sandmann, Numerical solution of level dependent quasi-birth-and-death processes, *Proc. Comput. Sci.*, **1** (2010), 1561–1569. https://doi.org/10.1016/j.procs.2010.04.175

45. M. Xi, W. Sun, J. Chen, Survey of derivative-free optimization, *Numer. Algebra Control Optim.*, **10** (2020), 537–555. https://doi.org/10.3934/naco.2020050

46. J. Larson, M. Menickelly, S. M. Wild, Derivative-free optimization methods, *Acta Numer.*, **28** (2019), 287–404. https://doi.org/10.1017/S0962492919000060

47. A. Dudin, S. Dudin, A. Melikov, O. Dudina, Framework for analysis of queueing systems with correlated arrival processes and simultaneous service of a restricted number of customers in scenarios with an infinite buffer and retrials, *Algorithms*, **17** (2024), 493. https://doi.org/10.3390/a17110493