AIMS *Mathematics*

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

AIMS Mathematics, 9(12): 33365–33385.
DOI: 10.3934/math.20241592
Received: 17 October 2024
Revised: 05 November 2024
Accepted: 11 November 2024
Published: 25 November 2024

Research article

Analysis of a multi-server retrial queue with a varying finite number of sources

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; Tel: +39-3381838978.

Abstract: A multi-server retrial queue with a finite number of sources of requests was considered. In contrast to similar models studied in the literature, we assumed this number is not constant but changes its value in a finite range. During the stay in the system, each source generates the service requests. These requests are processed in a finite pool of servers. After service completion of a request, the source is granted the possibility to generate another request. If the source does not use this possibility during an exponentially distributed time, it is deleted from the system. If the request finds all servers busy, it can make repeated attempts to enter the service. If all servers are busy, the request may depart from the system without service. In this case, with a fixed probability, the source that generated this request is deleted from the system. Sources arrive according to a Markov arrival process. If the number of sources in the system at the arrival epoch has the maximum allowed number, the arriving source is lost. This system is a more adequate model of many real-world systems than the standard finite source queue. Analysis of the considered system required a four-dimensional continuous-time Markov chain. The generator of the chain was obtained as a block matrix with four levels of nesting. The stationary distribution of this Markov chain was found numerically as well as the values of the system's performance measures. The dependence of these measures on the maximum allowed number of sources and the number of servers was numerically clarified. An example of solving an optimization problem was presented.

Keywords: finite-source queueing model; Markov arrival process; retrial; multidimensional Markov chains; performance; modeling
Mathematics Subject Classification: 60K25, 60K30, 60M20

1. Introduction

The phenomenon of request retrials, when the request cannot be admitted for service immediately upon arrival, is inherent in a variety of real-world systems. Examples of such systems and descriptions of the state-of-the-art methods of analysis of retrial queues can be found in the monographs [1–8].

Due to the wide applicability of retrial queues for analysis of various communication networks (including wireless networks) and contact centers, these queues are a popular subject of research despite the essential mathematical difficulties of their study compared to the queues with losses and buffers. These difficulties are caused by the state-inhomogeneous behavior and higher dimension of the stochastic processes describing such systems.

The overwhelming majority of the existing literature in the field of retrial queues is devoted to systems where the number of retrying customers is not restricted (systems with an infinite source of customers). At the same time, the retrial queues with a finite source also have received attention. Surveys of early works about the retrial queues with a finite source can be found in [9] and [10]. The majority of these works, including [9], consider single-server queues. As early works, the papers [11, 12] about the single-server queues can be mentioned.

In the paper [12] by A.G. de Kok, the author considers the system with an arrival rate depending on the number of customers in orbit (those who met the busy server and will make repeated attempts to enter service) and the state (busy or idle) of the server. Three special cases are considered there: systems with infinite orbit capacity, systems with a finite orbit capacity, and systems with quasi-random input. In the latter case, it is assumed that there are N identical permanently active sources. Every source can generate service requests. If a source generates a request when the server is idle, this request is immediately admitted for service; otherwise, the request joins the orbit. When the service of the request is completed, the source becomes free and sends out a new request after an exponentially distributed time. A non-free source, i.e., a source whose requests are in service or orbit, cannot generate a new request.

Namely, the latter special case of the request admission and retrial management is called in the literature the retrial system with a finite number of sources of calls. As the first paper devoted to the analysis of multi-server retrial queues with a finite number of sources of requests, the paper [10] by G. I. Falin can be mentioned. The $M/M/c$-type queue is considered there. According to [10], each source can be in one of three states: a request generated by this source receives service; the source is sending repeated calls (i.e., waiting for service in the orbit); the source is free and will generate the request in an exponentially distributed time with a constant rate.

From the point of view of potential real-world applications, the source of the request can be interpreted as the pair of equipment (access point or automated workplace) and a human (operator) who uses this equipment for uploading or downloading information via communication channels (servers). An automated workstation (AWP) is a system in which all the tools and programs necessary for work are combined into one environment, often under the control of specialized software. An AWP is designed to increase the efficiency of an employee, provide easy access to the necessary resources, and centralize all business processes. The mentioned assumption that the number of sources is permanent does not sound realistic because some of the existing access points can be temporarily unavailable (broken, under maintenance, etc.), and the free operators could have different activities at different periods (morning, day, evening and night), and generate the sources of requests only during working hours or during night.

The source of the requests can also be interpreted as a person having, e.g., a membership card, season ticket, subscription, etc., who has to be mandatorily provided service in a system if the required service equipment is currently available. Each source residing in the system can have a rest between the required sequential services or can receive service, or they can wait for the release of some server if all servers are occupied. In such systems, there is a fixed maximum value of the number of sources that can stay in the system at the same time and generate requests, while the actual number of residing sources is variable. Along with the already mentioned examples with the user having access to the system resource via certain access equipment, e.g., an automatic workplace or a computer having secure access to some specific data, the considered model can be applied for the description of the operation of a variety of real-world systems (communication, transportation, entertainment, banking, etc.) where the pool of users that can be serviced simultaneously is limited by a certain number. This number has to be properly chosen at the stage of the system design to balance the number of available service devices and size of the pool of potential users in such a way to achieve the maximal throughput of the system and, at the same time, to avoid overcrowding and provide a satisfactory level of service of the users admitted to the pool.

The main contribution of this paper is that we analyze the retrial queueing model where the number of sources is variable. The maximum number of sources that can reside in the system is fixed and is set to be equal to a finite integer number M. Therefore, any number $m \in \{0, 1, 2, \dots, M\}$ of sources can stay in the system (be active) at an arbitrary moment. After an exponentially distributed time, the free source, which did not generate a new request, leaves the system permanently. New sources arrive in the system and are admitted if the number of sources already processed in the system is less than M and are rejected otherwise. The proper choice of the optimal number of M and the number of servers in the system that guarantee the maximal throughput of the system and satisfaction of the fixed requirements for the quality of service is the non-trivial and challenging problem. The results of this paper provide the tool for solving this problem.

In the terminology common in the literature devoted to an analysis of IEEE 802.11 DCF protocol as a random channel access scheme based on CSMA/CA, the standard retrial queueing model with a finite number of sources corresponds to the saturated system where each station always has a packet for transmission, see [13]. Here we consider the non-saturated queueing model with a finite number of sources. As was mentioned in many papers, e.g., in [14], non-saturated systems better describe real-world systems.

For reasons of mathematical generality, in this paper, we assume that source arrival is described not by the stationary Poisson process, as in the majority of existing papers, but by a quite general Markov arrival process (MAP), see, e.g., [15–18]. The MAP is a well-recognized model of correlated bursty flows in modern telecommunication networks and contact centers as well as other real-world systems, see, e.g., [19].

The model considered in this paper has certain common features with the model analyzed in [20]. In the terminology of our present paper, the model considered in [20] assumes a variable number of active sources. New sources arrive according to a MAP. Request generation by a source is terminated after a geometrically distributed number of requests is already generated. The difference is that here we assume that a new request cannot be generated by a source until the previous request, which was generated by this source, completes its service. In [20], it is assumed that the source can generate many requests in turn without waiting for their service beginning and completion. Our model is more complicated for study than the one analyzed in [20] because the successive intervals between generations of requests are not independent, identically distributed random variables but have a complicated structure and include possible inter-retrial times.

A quite general multi-server retrial queue with a fixed finite number of sources of requests was considered in [21]. Time until the next request generation by a source after service completion of the previous request generated by this source does not have an exponential but more general so-called phase type (PH) distribution; for the definition and properties, see, e.g., [22, 23]. In contrast to our model, the number of sources in [21] is permanent (i.e., the system is saturated). Note that the additional features of the model considered in [21] are the presence of negative customers and the influence of the external random environment.

The multi-server retrial queues with a fixed, finite number of sources of requests were also considered in the following papers. In [24], a retrial queue with the search for balking and impatient customers from the orbit was considered. In [25], it was assumed that service time had a phase-type distribution. Two known ways for monitoring the phases of service time in all busy servers called TPFS (track phase for server) and CSFP (count server for phase) were considered. For more details about these ways, see, e.g., [26]. In [27], a finite-source queueing system consisting of heterogeneous servers with unequal service intensities was under study. In [28], a finite-source retrial queue with non-reliable heterogeneous servers was investigated via computer simulation.

Note that the retrial systems with an infinite source of arrivals defined by the MAP and phase type distribution of the service time have been intensively studied in the existing literature, see, e.g., [29–34].

A brief outline of the paper is as follows. The mathematical model under study is described in detail in Section 2. Section 3 contains a description of the process of changing the states of the system, which is essentially more complicated than the corresponding process for the system with a fixed number of sources, and briefly touches on the problem of the computation of the stationary distribution of this process. Formulas for the computation of the main performance measures of the system given the known stationary distribution of the system states are the main object of Section 4. Numerical results are given in Section 5. Section 6 briefly summarizes the results of the presented analysis and touches on possible directions for further research.

2. Mathematical model

Let us consider a multi-server retrial queue, the scheme of operation of which is presented in Figure 1.

A MAP-flow of sources of requests enters the system. This input flow is specified by the control process ν_t, $t \geq 0$, which is an irreducible continuous-time Markov chain (MC). This chain has a finite state space $\{1, 2, \ldots, W\}$ and the generator D of size W that can be represented as the sum of two matrices, D_0 and D_1. The entries of the non-negative matrix D_1 define the rates of transitions of the MC ν_t at which the sources are generated. The matrix D_0 is the sub-generator. Its diagonal entries are negative. The modules of these entries define the departure rates of the MC ν_t from its states. The non-diagonal entries of matrix D_0 are non-negative and define the rates of the MC ν_t transitions within its state space at which sources do not arrive. The average intensity of source arrival is denoted as λ and is calculated as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_W)$ is the invariant probability vector of the MC ν_t.

It is defined as the unique solution of the system $\boldsymbol{\theta} D = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Here and throughout the paper, $\mathbf{e}$ is a column vector of suitable size consisting of ones, and $\mathbf{0}$ is a row vector of suitable size consisting of zeros. A more detailed description of the MAP and formulas for determining its characteristics, including the correlation and variation coefficients, can be found in [15–18].

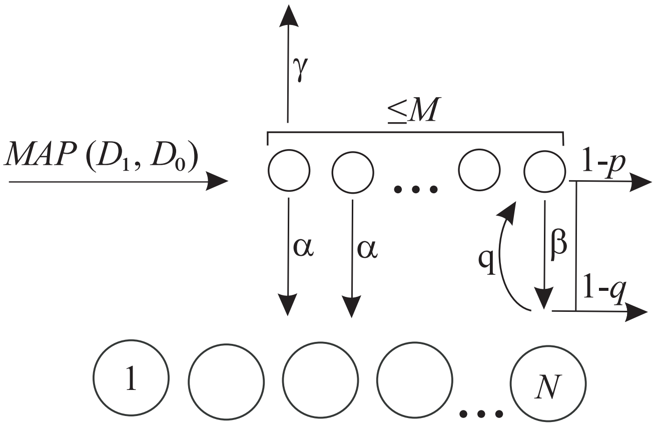

Figure 1. Scheme of the system operation.

It is assumed that there can be no more than M sources processed in the system at the same time. If M sources are already presented in the system at an arrival moment, the arrived source is permanently lost. Otherwise, it starts processing in the system as follows.

The number of independent identical servers that provide service to requests is equal to N. The admitted source, independently of other sources residing in the system, generates a request for service in exponentially distributed time with intensity α, $\alpha > 0$. If there is an idle server at the request generation epoch, then this request begins service at any of the idle servers. The service time is exponentially distributed with the parameter μ. Until the service of the request is finished, the source that has generated this request resides in the system but cannot generate new requests.

If, at a request generation moment, all servers are busy, then with probability q, $0 < q \leq 1$, the source decides to make repeated attempts to manage the service of the generated request. With the complementary probability $1 - q$, the request is lost. The time between the repeated attempts to send the request to service has an exponential distribution with parameter β, $\beta > 0$. If the attempt was successful, i.e., some server is available at the retrial epoch, then the request begins service. If the attempt to catch an idle server is unsuccessful, the source decides to continue attempts with probability q, and with the complementary probability, the request is lost. If a request leaves the system without being serviced, then its source remains in the system with probability p, and with complementary probability $1 - p$, this source leaves the system forever. The source that generated this request may generate a new request after the exponentially distributed time with the intensity α, $\alpha > 0$. Thus, any source residing in the system can be in the following three states: (i) blocked, when its request receives service by one of the servers; (ii) making repeated attempts; (iii) free (thinking before generating a new request). Note that a newly admitted source always turns to the free state.

We assume that the time of the source's stay in the system is limited. Namely, a free source leaves the system after an exponentially distributed time with intensity γ, $\gamma > 0$, if it did not generate a request during this time. By default, it is assumed that γ is much less than α, i.e., a source can generate, with a non-negligible probability, at least several requests during its residence in the system.

3. The process of changing the states of a system and its stationary distribution

Let i_t be the number of sources in the system, $i_t = \overline{0,M}$; m_t be the number of free sources, $m_t = \overline{0,i_t}$; n_t be the number of busy servers, $n_t = \overline{0,\min\{i_t - m_t, N\}}$; and ν_t be the state of the control process of the MAP, $\nu_t = \overline{1,W}$, at the moment t, $t \geq 0$. Here, notation like $i_t = \overline{0,M}$ means that i_t admits values from the set $\{0, 1, \dots, M\}$.

It can be verified that the random process $\xi_t = \{i_t, m_t, n_t, \nu_t\}$, $t \geq 0$, is an irreducible continuous-time MC.

To facilitate the work with a four-dimensional MC ξ_t, let us enumerate its states in the lexicographic order. We call the set of the states, which have the value (i, m) of the components $\{i_t, m_t\}$, sub-level (i, m), $m = \overline{0,i}$, $i = \overline{0,M}$. The set $((i, 0), (i, 1), \dots, (i, i))$ is called level i, $i = \overline{0,M}$.

We will use the following denotations:

The square matrix $\hat{I}$ has size $N + 1$ and is defined by $\hat{I} = \mathrm{diag}\{0, 0, \dots, 0, 1\}$, i.e., it is the diagonal matrix with all zero diagonal entries except the last entry that is equal to 1; I_W is the identity matrix of size W; and $\delta_{i,j}$ is Kronecker's delta (equal to 1 if $i = j$ and 0, otherwise).

The following statement is true.

Theorem 1. *The generator Q of the MC ξ_t, $t \geq 0$, has a tri-block-diagonal structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & \dots & O & O & O \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \dots & O & O & O \\ O & Q_{2,1} & Q_{2,2} & \dots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O & O & O & \dots & Q_{M-1,M-2} & Q_{M-1,M-1} & Q_{M-1,M} \\ O & O & O & \dots & O & Q_{M,M-1} & Q_{M,M} \end{pmatrix} \qquad (1)$$

where:

(1) The diagonal blocks $Q_{i,i}$, $i = \overline{0,M}$, have a tri-block-diagonal structure with non-zero diagonal blocks $Q_{i,i}^{m,m}$, $m = \overline{0,i}$, the updiagonal blocks $Q_{i,i}^{m,m+1}$, $m = \overline{0,i-1}$, and the sub-diagonal blocks $Q_{i,i}^{m,m-1}$, $m = \overline{1,i}$, defined as:

- *blocks $Q_{i,i}^{m,m}$ are two-block diagonal matrices with the diagonal blocks*

$$(Q_{i,i}^{m,m})^{n,n} = D_0 - \Big(m\alpha + (i - m - n)\beta + m\gamma + n\mu\Big)I_W$$

$$+\delta_{n,N}\Big(q(i - m - n)\beta + p(1 - q)m\alpha\Big)I_W + \delta_{i,M}D_1,\ n = \overline{0,\min\{i - m, N\}}, \qquad (2)$$

and the updiagonal blocks

$$(Q_{i,i}^{m,m})^{n,n+1} = (i - m - n)\beta I_W,\ n = \overline{0,\min\{i - m, N\} - 1}. \qquad (3)$$

- *The blocks $Q_{i,i}^{m,m+1}$ for $i-m \le N$ have the form*

$$Q_{i,i}^{m,m+1} = \begin{pmatrix} O & O & O & \dots & O \\ (Q_{i,i}^{m,m+1})^{1,0} & O & O & \dots & O \\ O & (Q_{i,i}^{m,m+1})^{2,1} & O & \dots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \dots & (Q_{i,i}^{m,m+1})^{i-m,i-m-1} \end{pmatrix}, \tag{4}$$

where

$$(Q_{i,i}^{m,m+1})^{n,n-1} = n\mu I_W,\ n = \overline{1, i-m}. \tag{5}$$

For $i-m > N$, the blocks $Q_{i,i}^{m,m+1}$ have non-zero subdiagonal blocks

$$(Q_{i,i}^{m,m+1})^{n,n-1} = n\mu I_W,\ n = \overline{1, N},$$

and one non-zero last diagonal block

$$(Q_{i,i}^{m,m+1})^{N,N} = p(1-q)(i-m-N)\beta I_W. \tag{6}$$

- *The blocks $Q_{i,i}^{m,m-1}$ for $i-m < N$ have the form*

$$Q_{i,i}^{m,m-1} = \begin{pmatrix} O & (Q_{i,i}^{m,m-1})^{0,1} & O & \dots & O \\ O & O & (Q_{i,i}^{m,m-1})^{1,2} & \dots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \dots & (Q_{i,i}^{m,m-1})^{i-m,i-m+1} \end{pmatrix}, \tag{7}$$

where

$$(Q_{i,i}^{m,m-1})^{n,n+1} = m\alpha I_W,\ n = \overline{0, i-m}.$$

For $i-m \ge N$, the blocks $Q_{i,i}^{m,m-1}$ have non-zero updiagonal blocks

$$(Q_{i,i}^{m,m-1})^{n,n+1} = m\alpha I_W,\ n = \overline{0, N-1},$$

and one non-zero last diagonal block

$$(Q_{i,i}^{m,m-1})^{N,N} = qm\alpha I_W. \tag{8}$$

(2) *The updiagonal blocks $Q_{i,i+1}$, $i = \overline{0, M-1}$, have the form*

$$Q_{i,i+1} = \begin{pmatrix} Q_{i,i+1}^{0,0} & Q_{i,i+1}^{0,1} & O & \dots & O & O \\ O & Q_{i,i+1}^{1,1} & Q_{i,i+1}^{1,2} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ O & O & O & \dots & Q_{i,i+1}^{i,i} & Q_{i,i+1}^{i,i+1} \end{pmatrix}, \tag{9}$$

where $Q_{i,i+1}^{m,m}$, $m = \overline{0, i}$, are zero matrices of size $(\min\{i-m, N\}+1)W \times (\min\{i+1-m, N\}+1)W$, *and $Q_{i,i+1}^{m,m+1}$, $m = \overline{0, i}$, are the block-diagonal matrices with the diagonal blocks $(Q_{i,i+1}^{m,m+1})^{n,n} = D_1$, $n = \overline{0, \min\{i-m, N\}}$.*

(3) The subdiagonal blocks $Q_{i,i-1}$, $i = \overline{1,M}$, have the form

$$Q_{i,i-1} = \begin{pmatrix} Q_{i,i-1}^{0,0} & O & O & \dots & O & O \\ Q_{i,i-1}^{1,0} & Q_{i,i-1}^{1,1} & O & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & Q_{i,i-1}^{i-1,i-2} & Q_{i,i-1}^{i-1,i-1} \\ O & O & O & \dots & O & Q_{i,i-1}^{i,i-1} \end{pmatrix}, \tag{10}$$

where the subdiagonal blocks are defined as $Q_{i,i-1}^{m,m-1} = \gamma m I_{\min\{i-m,N\}+1} \otimes I_W$, $m = \overline{1,i}$, if $i-m < N$, and $Q_{i,i-1}^{m,m-1} = (\gamma m I_{N+1} + \alpha m(1-q)(1-p)\hat{I}) \otimes I_W$, if $m = \overline{1,i}$. The diagonal blocks $Q_{i,i-1}^{m,m}$, $m = \overline{0,i-1}$, are zero matrices of size $(\min\{i-m,N\}+1)W \times (\min\{i-1-m,N\}+1)W$, *when $i - m \leq N$. For $i - m > N$, they are matrices of size* $(N+1)W \times (N+1)W$, *with all zero blocks except the last diagonal block, $(Q_{i,i-1}^{m,m})^{N,N} = (1-q)(1-p)(i-m-N)\beta I_W$.*

Proof of Theorem 1 is implemented via the careful analysis of all possible transitions of the four-dimensional MC $\xi_t = \{i_t, m_t, n_t, \nu_t\}$ during an infinitesimal time. The generator Q is a block matrix whose blocks $Q_{i,j}$ define transition rates from the states that belong to the level i to the states that belong to the level j. The blocks $Q_{i,j}$ in turn consist of the sub-blocks $Q_{i,j}^{m,m'}$ defining transition rates from the states that belong to the sub-level (i,m) to the states that belong to the sub-level (j,m'). The sub-blocks $Q_{i,j}^{m,m'}$ consist of block matrices $(Q_{i,j}^{m,m'})^{n,n'}$ whose entries are the matrices that define transition rates of the components $\{i_t, m_t, n_t\}$ of the MC ξ_t from the state (i,m,n) to the state (j,m',n') along with possible transitions of the component ν_t of the chain ξ_t.

Due to the properties of the exponential distribution of time until the generation of a new request by a free source, inter-retrial times, request service times, and time until the deletion of the inactive free source, each of the components $\{i_t, m_t, n_t\}$, during the time interval of the infinitesimal length, can keep its value or increase or decrease it by 1.

This implies that all the listed matrices, $Q_{i,j}$, $Q_{i,j}^{m,m'}$, and $(Q_{i,j}^{m,m'})^{n,n'}$ have all zero blocks except, probably, the diagonal, updiagonal, and subdiagonal blocks. For the same reason, the generator Q is a tri-block diagonal structure (1) because the number of sources in the system can maintain its value or increase it by 1, if a new source arrives and is admitted to the system, or decrease it by 1, if a source leaves the system due to too long of a stay in the free state or the loss of a request generated by this source.

The matrices $Q_{i,i}$, $i = \overline{0,M}$, also have the tri-block diagonal structure because, under the fixed number i of sources residing in the system, the number of free sources m can retain its value or increase or decrease it by 1. Let us comment on the structure of the blocks $Q_{i,i}^{m,m'}$, $m' = m-1, m, m+1$, of the matrix $Q_{i,i}$.

First of all, consider the expression for the block $Q_{i,i}^{m,m}$. This block is not a tri-block diagonal, but a two-block diagonal matrix. This is because, under the fixed value (i,m) of the components $\{i_t, m_t\}$ of the MC ξ_t, the component n_t cannot decrease its value. Such a decrease corresponds to the service completion of some requests on one busy server. However, this implies that the source that has generated this request becomes idle, and the number of free servers must increase to $m+1$ and not remain equal to m. Therefore, $(Q_{i,i}^{m,m})^{n,n-1} = O$. The form (2) of the block $(Q_{i,i}^{m,m})^{n,n}$ is explained as follows.

The diagonal entries of this block are negative. The module of each entry is equal to the exit rate of the MC ξ_t from the corresponding state. Such an exit can occur due to the exit from its states of the tri-dimensional process $\{i_t, m_t, n_t\}$ or of the process ν_t. Note that, during the interval of an infinitesimal length, only one of the processes $\{i_t, m_t, n_t\}$ and ν_t can exit from its states. The other one remains in its state. The rate of the exit of the processes $\{i_t, m_t, n_t\}$ for the fixed number i of sources presenting in the system, m of free servers, and n of busy servers, $n = \overline{0, N-1}$, is equal to $m\alpha + (i-m-n)\beta + m\gamma + n\mu$. This is true because each of m free sources can generate a new request (with rate α) or depart from the system (with rate γ), each busy server can complete service (with rate μ), and each retrying request (the number of such requests is equal to $(i-m-n)$) can retry (with rate β). When $n = N$, the request generated by each of m free sources is lost with probability $(1-q)$ and with probability p the corresponding source remains in the system. This explains the presence of the additional summand $\delta_{n,N}\Big(q(i-m-n)\beta + p(1-q)m\alpha\Big)I_W$ in the right-hand side of (2).

The multiplier I_W here reflects the mentioned-above fact that the underlying process ν_t of the MAP should not make any change during the small interval of time if some changes in the value of the process $\{i_t, m_t, n_t\}$ occur. A transition of the process $\{i_t, m_t, n_t, \nu_t\}$ without changing the value (i, m, n) of the components $\{i_t, m_t, n_t\}$ can also occur due to the transitions of the process ν_t defined by the matrix D_0. The diagonal entries of this matrix are negative, and the module of each such entry defines the exit rate from the corresponding state. The non-diagonal entries of this matrix define transition rates without the generation of a new source. This explains the presence of the summand D_0 in the right-hand side of formula (2). The presence of the summand $\delta_{i,M}D_1$ here is explained by the fact that transitions of the process ν_t, the rates of which are given by the matrix D_1, do not lead to any change of the values of the components $\{i_t, m_t, n_t\}$ of the MC ξ_t because the arriving source is not admitted to the system and is lost when $i = M$. Thus, we have completely explained the form (2) of the blocks $(Q_{i,i}^{m,m})^{n,n}$.

The form (3) of the block $(Q_{i,i}^{m,m})^{n,n+1}$ is clear because the transition of the process $\{i_t, m_t, n_t\}$ from the state (i, m, n) to the state $(i, m, n+1)$ can happen due to the successful retrial by one of $(i-m-n)$ retrying with rate β sources, $n = \overline{0, \min\{i-m, N\}-1}$.

Let us now explain the form (4) of the block $Q_{i,i}^{m,m+1}$ when $i-m \le N$. If $i-m \le N$, then $\min\{i-m, N\} = i-m$ and the number n of busy servers admits values from 0 to $i-m$. When the number of free servers increases from m to $m+1$, the maximum value of n reduces to $i-m-1$. Therefore the matrix $Q_{i,i}^{m,m+1}$ with blocks $(Q_{i,i}^{m,m+1})^{n,n'}$ is not square. It has $i-m+1$ block rows and $i-m$ columns. Among these blocks, only the blocks $(Q_{i,i}^{m,m+1})^{n,n-1}$, $n = \overline{1, i-m}$, are non-zero. This is because the increase of the number of free sources from m to $m+1$ happens only at the moment of service completion in one of n busy servers (the rate of service completions is equal to $n\mu$). This proves formula (5).

If $i-m > N$, then $\min\{i-m, N\} = N$. Therefore, the matrix $Q_{i,i}^{m,m+1}$ is square. It has non-zero subdiagonal entries equal to $n\mu I_W$, $n = \overline{1, N}$, and one diagonal block given by formula (6). This formula is obvious because the transition of the components $\{i_t, m_t, n_t\}$ from the state (i, m, N) to the state $(i, m+1, N)$ happens when all servers are busy and retrial occurs (with rate $(i-m-N)\beta$). This retrial is not successful. According to the description of the model, with probability $(1-q)$, the retrying request is lost, but with probability p, the source generating this request does not leave the system but becomes free. Thus, the number of free sources becomes equal to $m+1$.

Let us now prove formula (7) for the block $Q_{i,i}^{m,m-1}$. The transition from the sub-level (i, m) to the sub-level $(i, m-1)$ can occur when one of m free sources generates a request (with rate $m\alpha$). Because in the considered case $i - m < N$, the generated request is accepted for service and the component n_t of the MC ξ_t increases by one. Thus, the non-square matrix $Q_{i,i}^{m,m-1}$ has non-zero only blocks $(Q_{i,i}^{m,m-1})^{n,n+1}$, which means that this matrix has structure (7) with the blocks $(Q_{i,i}^{m,m-1})^{n,n+1}$ equal to $m\alpha I_W$, $n = \overline{0, i-m}$.

If $i - m \geq N$, then the square matrix $Q_{i,i}^{m,m-1}$ has, along with the same updiagonal blocks, the non-zero matrix $(Q_{i,i}^{m,m-1})^{N,N}$ reflecting the scenario when one of m sources generates a request, it is lost because all servers are busy and the source, which has generated this request, does not leave the system but becomes the source generating the retrials. As a result, we obtain formula (8).

Let us now explain form (9) of the matrices $Q_{i,i+1}$, $i = \overline{0, M-1}$. These matrices consist of the blocks $Q_{i,i+1}^{m,m'}$ where $m = \overline{0,i}$, $m' = \overline{0, i+1}$. Thus, these matrices are not square. They have $i+1$ block rows and $i+2$ block columns. Because the transition from level i to level $i+1$ occurs via a new source arrival (the rates of transition of the component v_t of the MC ξ_t are given by the matrix D_1) and this arrival implies the increase by 1 of the number of free sources, we obtain the formulas given in the formulation of Theorem 1 for the computation of the matrix $Q_{i,i+1}$ and its blocks. Note that, as is stated in Theorem 1, the blocks $Q_{i,i+1}^{m,m}$ are zero blocks for $m = \overline{0,i}$ and we could directly replace these blocks with zero matrices. We preferred to keep these matrices in (9) to give exact information about the size of the corresponding zero matrices.

Expression (10) for the blocks $Q_{i,i-1}$, $i = \overline{1, M}$, having $i+1$ block rows and i block columns, and their sub-blocks, is explained similarly. The decrease in the number of residing sources from i to $i-1$ may not imply the change in the number of free sources (if the decrease occurs due to the source departure caused by the unsuccessful retrial when all servers are busy). This explains the presence of the blocks $Q_{i,i-1}^{m,m}$, $m = \overline{0, i-1}$, in the generator. The decrease in the number of residing sources from i to $i-1$ may also cause a decrease in the number of free servers if the source departs from the system due to a long stay without a request generation or due to the departure of the free source caused by the generated request rejection. This explains the presence of non-zero blocks $Q_{i,i-1}^{m,m-1}$ in the generator.

We have completed the explanation of the form of the generator, its blocks, and sub-blocks. Theorem 1 is proven.

Remark 1. *It was mentioned above that the main contribution of this paper is the possibility of variation in the number of sources in the system. This number, if it does not have a maximum allowed value M, increases by one when a new source arrives in the MAP. This number, if it is positive, decreases by one when a free source departs from the system due to too long of a period without a request generation or rejection of the request generated by this source. The difficulty of the analysis of the system with a varying number of sources compared to the classical system with a permanent number of sources is easily explained by the fact that here we need to analyze the behavior of the* **four**-*dimensional MC ξ_t while the classical system, see [10], is analyzed via the consideration of the* **two**-*dimensional MC $\{m_t, n_t\}$, $m_t = \overline{0, M}$, $n_t = \overline{0, N}$. This leads to an essentially more complicated form of the generator describing the behavior of the system (four levels of blocks nesting instead of only two) and an essentially larger size of the blocks, which is essential for the computer implementation of the computation of the steady-state distribution of the system states under the realistic values of the number of servers N and maximum number M of sources that can be processed simultaneously.*

It is easy to verify that the MC ξ_t is an irreducible one and has a finite state space. Therefore, the positive limits

$$\pi(i,m,n,\nu),\ i=\overline{0,M},\ m=\overline{0,i},\ n=\overline{0,\min\{i-m,N\}},\ \nu=\overline{1,W},$$

called stationary probabilities of the states of the MC ξ_t, exist for any values of the system parameters. According to the introduced lexicographic ordering of the states of the MC, let us form the row vectors $\boldsymbol{\pi}(i,m)$ of the stationary probabilities of the states that belong to the sub-level (i,m), $i=\overline{0,M}$, $m=\overline{0,i}$, and the row vectors $\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0),\boldsymbol{\pi}(i,1),\dots,\boldsymbol{\pi}(i,i))$, $i=\overline{0,M}$, of the stationary probabilities of the states that belong to the level i, $i \geq 0$.

It is well known that the vectors $\boldsymbol{\pi}_i$, $i=\overline{0,M}$, can be found as the unique solution to the system of the linear algebraic equations

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

where Q is the generator of the MC ξ_t defined by Theorem 1. Although this system has a finite number of equations, this number can be large. Thus, the tri-block diagonal structure of the generator has to be effectively taken into account to solve it. Algorithms from [35, 36] can be recommended for this goal.

4. Performance characteristics of the system

Having calculated the vectors of stationary probabilities of the system states, we can calculate their main stationary performance characteristics. Expressions for calculating some of them are given below.

The average number of sources in the system is calculated using the formula

$$L_{source}=\sum_{i=1}^{M} i\boldsymbol{\pi}_i\mathbf{e}.$$

The average number of free sources in the system is given by

$$L_{free-source}=\sum_{i=1}^{M}\sum_{m=1}^{i} m\boldsymbol{\pi}(i,m)\mathbf{e}.$$

The average number of occupied servers is equal to

$$L_{blocked-source}=\sum_{i=1}^{M}\sum_{m=0}^{i-1}\sum_{n=1}^{\min\{i-m,N\}} n\boldsymbol{\pi}(i,m,n)\mathbf{e}.$$

The average number of sources that repeat attempts is calculated using the formula

$$\begin{aligned}L_{retrial-source}&=\sum_{i=1}^{M}\sum_{m=0}^{i-1}\sum_{n=0}^{\min\{i-m-1,N\}}(i-m-n)\boldsymbol{\pi}(i,m,n)\mathbf{e}\\&=L_{source}-L_{free-source}-L_{blocked-source}.\end{aligned}$$

The average intensity of serviced requests is defined as

$$\lambda_{serv}=\mu L_{blocked-source}.$$

The probability of losing an arbitrary request is calculated as

$$P_{req-loss} = 1 - \frac{\lambda_{serv}}{\alpha L_{free-source}}.$$

The probability of losing an arbitrary source at the entrance to the system is calculated as

$$P_{source-loss-ent} = \frac{1}{\lambda}\sum_{m=0}^{M}\sum_{n=0}^{\min\{M-m,N\}} \boldsymbol{\pi}(M,m,n)D_1\mathbf{e}.$$

The probability of losing an arbitrary source since the request it generated did not immediately reach the service is calculated as

$$P_{source-loss-gen} = \frac{1}{\lambda}\sum_{i=N+1}^{M}\sum_{m=1}^{i-N}\alpha m(1-q)(1-p)\boldsymbol{\pi}(i,m,N)\mathbf{e}.$$

The probability of losing an arbitrary source due to an unsuccessful retrial to get service is given by

$$P_{source-loss-ret} = \frac{1}{\lambda}\sum_{i=N+1}^{M}\sum_{m=0}^{i-N}\beta(i-m-N)(1-q)(1-p)\boldsymbol{\pi}(i,m,N)\mathbf{e}.$$

The probability of an arbitrary source leaving the system due to long inactivity in the free state is defined as

$$\begin{aligned}P_{source-left} &= \frac{1}{\lambda}\sum_{i=1}^{M}\sum_{m=1}^{i} \gamma m\boldsymbol{\pi}(i,m)\mathbf{e}\\ &= 1 - P_{source-loss-ent} - P_{source-loss-gen} - P_{source-loss-ret}.\end{aligned}$$

The presence of two different formulas for the computation of the probability $P_{source-left}$ is helpful for the control of the accuracy of computation of the stationary probabilities of the system states.

The probability that at any given moment there is a free server in the system is calculated as

$$P_{free-server} = 1 - \sum_{i=N}^{M}\sum_{m=0}^{i-N}\boldsymbol{\pi}(i,m,N)\mathbf{e}.$$

5. Numerical example

The goals of the following numerical example are to demonstrate the feasibility of the proposed method for computing performance characteristics of the system and illustrate the dependencies of certain performance measures of the system on the number N of servers and the maximum number M of sources that can receive service in the system simultaneously. The possibility of solving an optimization problem is illustrated as well.

Let the MAP flow of sources arriving at the system be defined by the matrices

$$D_0 = \begin{pmatrix} -0.8 & 0 \\ 0 & -0.2 \end{pmatrix},\ D_1 = \begin{pmatrix} 0.75 & 0.05 \\ 0.01 & 0.19 \end{pmatrix}.$$

The fundamental rate λ of this flow is equal to 0.3. The coefficients of correlation and variation of inter-arrival times are $c_{cor} = 0.17$ and $c_{var} = 1.625$.

The rate of request generation by a free source is $\alpha = 0.2$. The rate of retrial generation is $\beta = 0.3$. The rate of the free source departure from the system is $\gamma = 0.01$. The service rate is $\mu = 1$. The probability that the source will not depart from the system when the loss of request generated by this source occurs is p $p = 0.8$. The probability that the source will make retrials if the request generated by this source meets all busy servers is $q = 0.9$. Let us vary the parameter N over the interval $[1, 10]$ with step 1, and the parameter M over the interval $[5, 100]$ also with step 1.

The dependence of the average number L_{source} of sources in the system on the parameters N and M is presented in Figure 2. The number L_{source} more or less quickly increases when the maximum number M of sources increases from 1 to about 60. After that, the increase becomes slow, and the average number L_{source} of sources stabilizes. The value L_{source} is slightly larger for a small number of servers N because when N is small, the source rarely becomes free and obtains a chance to depart from the system.

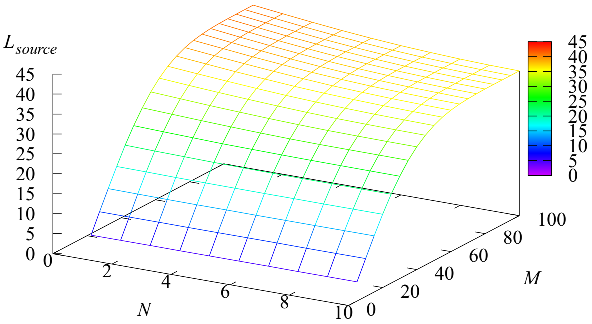

Figure 2. Dependence of the L_{source} on the parameters N and M.

The dependence of the average number $L_{blocked-source}$ of sources (the average number of busy servers) in the system on the parameters N and M is presented in Figure 3. This number is small when the maximum number M of sources and the total number of servers N are small and then increases when M and N grow. The increase caused by the increase of M becomes negligible for $M > 50$.

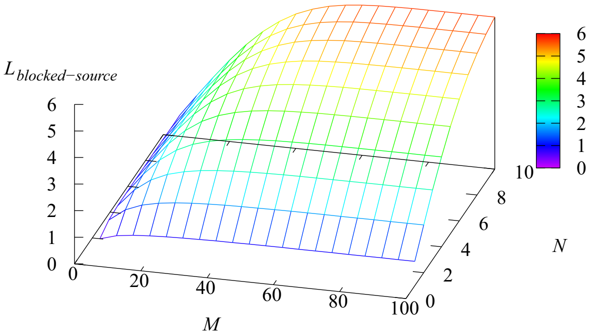

Figure 3. Dependence of the average number $L_{blocked-source}$ on the parameters N and M.

It is worth noting that because the service rate μ is equal to 1, Figure 3 also gives an illustration of the dependence of the average intensity λ_{serv} of serviced requests on the parameters N and M.

Figure 4 shows the dependence of the average number $L_{free-source}$ of free sources on the parameters N and M. This number increases with the growth of the maximum number M of sources and, especially, with the growth of the number of servers N.

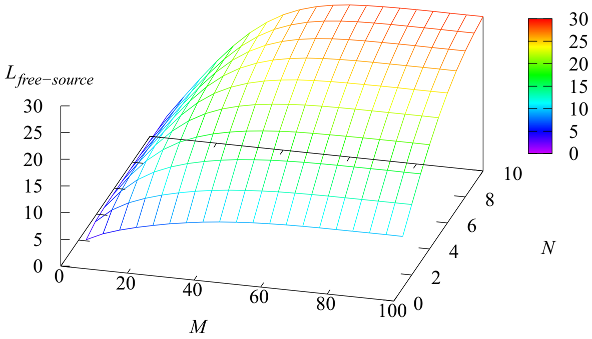

Figure 4. Dependence of the average number $L_{free-source}$ of free sources on the parameters N and M.

Figure 5 illustrates the behavior of the average number $L_{retrial-source}$ of sources that repeat attempts to send the generated request to service. This number is very large when the number N of servers is small while the number of competing sources is large (because the maximum number of admitted sources is large). $L_{retrial-source}$ becomes small when N is large and M is small.

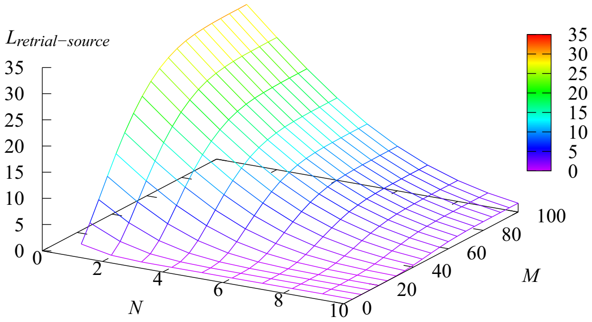

Figure 5. Dependence of the average number $L_{retrial-source}$ of sources that repeat attempts to send the generated request to service on the parameters N and M.

The behavior of the probability $P_{free-server}$ that at an arbitrary moment, there is a free server in the system is highlighted in Figure 6. This probability is small when M is large and N is small. When M is small and N is large, this probability is close to 1, which agrees with the intuitive reasoning.

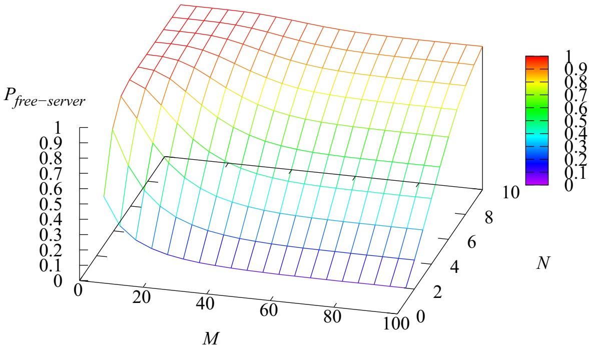

Figure 6. Dependence of the probability $P_{free-server}$ that at an arbitrary moment, there is a free server on the parameters N and M.

The surface giving the dependence of the loss probability $P_{req-loss}$ of an arbitrary request is presented in Figure 7. As anticipated, this probability is pretty large when M is large (and competition between the sources is high) and N is small. This probability quickly decreases when M decreases and N increases.

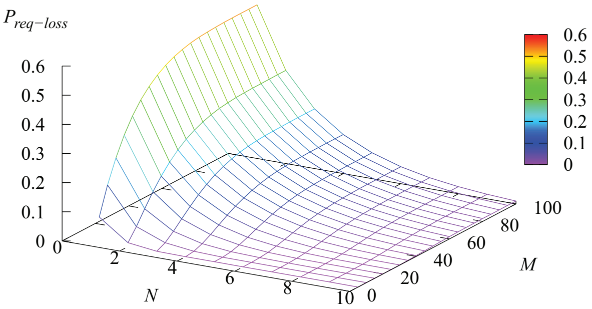

Figure 7. Dependence of the loss probability $P_{req-loss}$ of an arbitrary request on the parameters N and M.

The dynamic of the probability $P_{source-loss-ent}$ of losing an arbitrary source at the entrance to the system is shown in Figure 8. This probability is very high when only a few sources can work in the system together and essentially decreases when the maximum number M of sources that can operate in parallel increases. The influence of N is inessential.

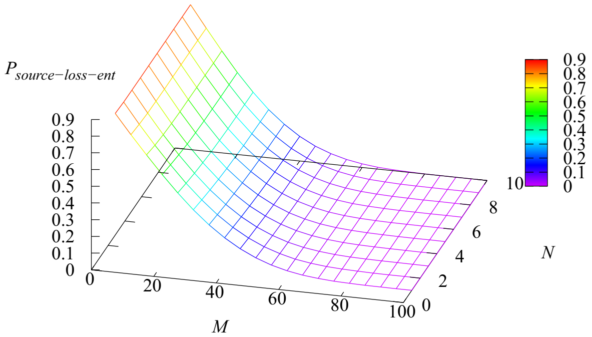

Figure 8. Dependence of the probability $P_{source-loss-ent}$ of losing an arbitrary source at the entrance to the system on the parameters N and M.

The dependence of the probability $P_{source-loss-gen}$ of losing an arbitrary source, because the request generated by this source did not immediately reach the service on the parameters N and M, is illustrated in Figure 9. The shape of the surface giving this dependence is less monotonous and intuitively clear than the majority of the surfaces presented above. This explains the motivation for the implemented study. Intuitive reasoning does not always help to understand the existing relations and make the proper managerial decisions. Mathematical, algorithmic, and numerical analysis is necessary to discover the existing dependencies under all combinations of the choice of system parameters.

Figure 10 shows the dependence of the probability $P_{source-loss-ret}$ of losing an arbitrary source due to an unsuccessful retrial to get service on the parameters N and M. As may be expected, this probability is very high when N is small and M is large, and essentially decreases when N increases and M decreases.

The probability $P_{source-left}$ that an arbitrary arriving source will leave the system due to long inactivity in the free state as the function of the parameters N and M is presented in Figure 11. This probability is high when N is large (and the system successfully provides service to almost all generated requests) and M is relatively large (greater than 50). When M is small, an essential part of the sources is lost at the entrance and, thus, has no chance to leave the system due to long inactivity in the free state.

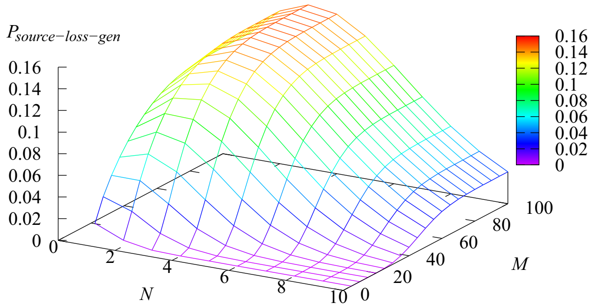

Figure 9. Dependence of the probability $P_{source-loss-gen}$ of losing an arbitrary source because the request generated by this source did not immediately reach the service on the parameters N and M.

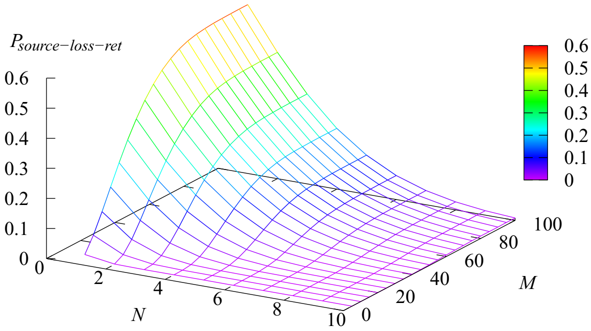

Figure 10. Dependence of the probability $P_{source-loss-ret}$ of losing an arbitrary source due to an unsuccessful retrial to get service on the parameters N and M.

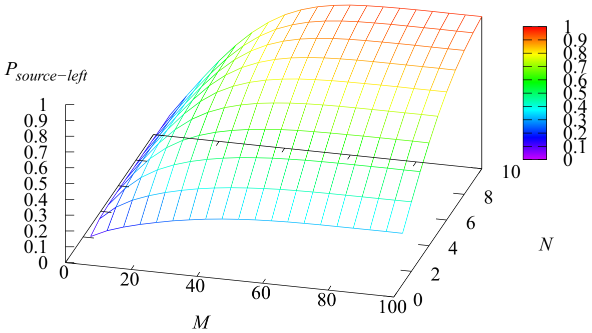

Figure 11. Dependence of the probability $P_{source-left}$ on the parameters N and M.

In the considered model, we have two parameters, N and M, that, as we see from the results of the numerical example, have an essential impact on the system performance. Therefore, various optimization problems arise. For example, if the number N of the servers is fixed, it is necessary to choose the maximum number M of sources (the number of automated workplaces), which can be processed in the system simultaneously, to provide the desired quality of service. Oppositely, if the maximum number M of sources is fixed, the number N of servers (or the respective service rates) has to be chosen. Otherwise, it is necessary to choose the optimal pair of parameters N and M.

The problem of fixing the proper criterion of quality of the system operation is important, but it can be solved, usually, based on common sense. For example, it seems reasonable to use the following criterion:

$$E(N, M) = a\lambda_{serv} - c_1\alpha L_{free-source}P_{req-loss} - c_2\lambda P_{source-loss-ent} \\ - c_3\lambda P_{source-loss-gen} - c_4\lambda P_{source-loss-ret} - dN$$

where

a is the profit gained by the system from the successful service of one request;

c_1 is the penalty for the loss of one request;

c_2 is the penalty for the loss of one source at the entrance to the system (due to the presence of the maximum allowed number of sources);

c_3 is the penalty for the loss of one source due to the loss of a request generated by this source;

c_4 is the penalty for the loss of one source due to the loss of a retrial generated by this source;

d is the payment for maintenance of each server during the unit of time; and

$E(N, M)$ is the average profit gained by the system per unit of time.

Let us fix the following values of the cost coefficients:

$$a = 1,\ c_1 = 3,\ c_2 = 2,\ c_3 = 15,\ c_4 = 20,\ d = 0.5.$$

Figure 12 shows the dependence of the cost criterion $E(N, M)$ on the parameters N and M.

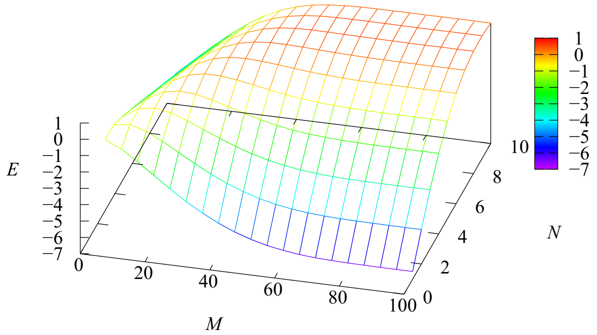

Figure 12. Dependence of the cost criterion $E(N, M)$ on the parameters N and M.

More information about the values of $E(N, M)$ for different values of N and M can be found in the table.

Table 1. Values of $E(N, M)$ for different values of N and M.

N	$E(N, 5)$	$M^*(N)$	$E(N, M^*(N))$	$E(N, 100)$
1	-0.811	5	-0.811	-6.535
2	-0.842	10	-0.689	-4.888
3	-1.234	15	-0.477	-3.384
4	-1.721	20	-0.229	-2.075
5	-2.22	25	0.02	-1.016
6	-2.72	35	0.283	-0.24
7	-3.22	45	0.463	0.253
8	-3.72	50	0.558	0.49
9	-4.22	60	0.529	0.513
10	-4.72	70	0.375	0.372

By $M^*(N)$ in the Nth row of the table, we mean the optimal value of M under the fixed value of N. Looking at the table, we can make the following inferences:

- If the maximum number M of sources is equal to 5, the value of the cost criterion is negative for any number N of servers.
- If the number N of servers is less than 5, the value of the cost criterion is negative for any maximum number M of sources.
- When N increases by 1, the optimal value of M increases by 5-10.
- When N increases up to the value $N = 8$, the value of $E(N, M^*(N))$ increases.
- A further increase in N causes the decrease of $E(N, M^*(N))$ because the charge for the use of an additional server exceeds the profit obtained due to the improvement of performance characteristics of the system.

The optimal values N^* and M^* of the parameters N and M that provide a maximum for the function $E(N, M)$ in this example are: $N^* = 8$, $M^* = 50$. The maximal value E^* of the cost criterion is equal to 0.558.

6. Conclusions

In this paper, we analyzed the essential generalization of a known retrial queue with a finite number of sources. Besides consideration of a multi-server queue, while a majority of existing papers focus on the single-server queue, we assume that the number of sources of requests (primary and retrials), which are processed in the system, is not constant but admits different values up to a finite maximum. Variation in the number of processed sources occurs due to deletion from the system of free sources with low activity and admission for service of new sources. The flow of the new sources is described by the MAP, which allows us to apply the analyzed model to a variety of real-world systems where flows of sources have bursty behavior.

As possible directions for further research, we can mention the systems with retrials of sources that arrive when the number of processed sources has the maximum allowed value, see, e.g., [37, 38]; systems with batch arrival of sources; systems with batch generation of requests by a source; systems with phase-type distribution of service time; and possible differences for requests that reach the server immediately upon generation or after retrials. For the latter systems, experience of [25] with the TPFS and CSFP methods for service phase tracking can be used. Consideration of different phase-type distributions of service time can be implemented with the use of the generalized phase-type distribution, see [39]. Similar to [21, 40], results can be extended also to the system operating in the random environment and to the semi-open networks with customer retrial in the case of the inner network overflow, see, e.g., [41].

Author contributions

Ciro D'Apice: Conceptualization, Investigation, Formal analysis, Writing-review and editing; Alexander Dudin: Conceptualization, Investigation, Formal analysis, Writing-original draft preparation, Writing-review and editing, Supervision; Sergey Dudin: Methodology, Software, Validation, Formal analysis, Writing-review and editing, Supervision; Rosanna Manzo: Methodology,

Software, Validation, Formal analysis, Writing-review and editing. All authors have read and approved the final version of the manuscript for publication.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. G. I. Falin, J. G. C. Templeton, *Retrial queues*, London: Chapman & Hall, 1997.
2. J. R. Artalejo, A. Gomez-Corral, *Retrial queueing systems*, Berlin: Springer, 2008.
3. G. Falin, A survey of retrial queues, *Queueing Syst.*, **7** (1990), 127–167. https://doi.org/10.1007/BF01158472
4. T. Yang, J. G. C. Templeton, A survey on retrial queues, *Queueing Syst.*, **2** (1987), 201–233. https://doi.org/10.1007/BF01158899
5. A. Gomez-Corral, A bibliographical guide to the analysis of retrial queues through matrix analytic techniques, *Ann. Oper. Res.*, **141** (2006), 163–191. https://doi.org/10.1007/s10479-006-5298-4
6. J. R. Artalejo, Accessible bibliography on retrial queues: progress in 2000–2009, *Math. Comput. Model.*, **51** (2010), 1071–1081. https://doi.org/10.1016/j.mcm.2009.12.011
7. J. Kim, B. Kim, A survey of retrial queueing systems, *Ann. Oper. Res.*, **247** (2016), 3–36. https://doi.org/10.1007/s10479-015-2038-7
8. C. S. Kim, S. H. Park, A. Dudin, V. Klimenok, G. Tsarenkov, Investigation of the $BMAP/G/1 \rightarrow / \bullet /PH/1/M$ tandem queue with retrials and losses, *Appl. Math. Model.*, **34** (2010), 2926–2940. https://doi.org/10.1016/j.apm.2010.01.003
9. G. Falin, J. R. Artalejo, A finite source retrial queue, *Eur. J. Oper. Res.*, **108** (1998), 409–424. https://doi.org/10.1016/S0377-2217(97)00170-7
10. G. Falin, A multiserver retrial queue with a finite number of sources of primary calls, *Math. Comput. Model.*, **30** (1999), 33–49. https://doi.org/10.1016/S0895-7177(99)00130-2
11. Y. N. Kornyshev, Design of a fully accessible switching system with repeated calls, *Telecommunications*, **23** (1969), 46–52.
12. A. G. Kok, Algorithmic methods for single server systems with repeated attempts, *Stat. Neerl.*, **38** (1984), 23–32. https://doi.org/10.1111/j.1467-9574.1984.tb01094.x
13. G. Bianchi, IEEE 802.11-saturation throughput analysis, *IEEE Commun. Lett.*, **2** (1998), 318–320. http://dx.doi.org/10.1109/4234.736171
14. Y. Lee, M. Y. Chung, T. J. Lee, Performance analysis of IEEE 802.11 DCF under nonsaturation condition, *Lect. Notes Comput. Sci.*, **17** (2008), 1–17. http://doi.org/10.1155/2008/574197
15. 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.
16. 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.

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

18. 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

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

20. S. A. Dudin, The $MAP/N/N$ retrial queueing system with time-phased batch arrivals, *Probl. Inform. Transmiss.*, **45** (2009), 270–281. https://doi.org/10.1134/S0032946009030089

21. J. Wu, Z. Liu, G. Yan, Analysis of the finite source $MAP/PH/N$ retrial G-queue operating in a random environment, *Appl. Math. Model.*, **35** (2011), 1184–1193. https://doi.org/10.1016/j.apm.2010.08.006

22. M. F. Neuts, *Matrix-geometric solutions in stochastic models: an algorithmic approach*, Chicago: Courier Corporation, 1994.

23. C. A. O'Cinneide, Phase-type distributions: open problems and a few properties, *Stochast. Models*, **15** (1999), 731–757. https://doi.org/10.1080/15326349908807560

24. P. Wüchner, J. Sztrik, H. de Meer, Finite-source $M/M/S$ retrial queue with search for balking and impatient customers from the orbit, *Comput. Networks*, **53** (2009), 1264–1273. https://doi.org/10.1016/j.comnet.2009.02.015

25. A. S. Alfa, K. S. Isotupa, An $M/PH/k$ retrial queue with finite number of sources, *Comput. Oper. Res.*, **31** (2004), 1455–1464. https://doi.org/10.1016/S0305-0548(03)00100-X

26. Q. M. He, A. S. Alfa, Space reduction for a class of multidimensional Markov chains: a summary and some applications, *INFORMS J. Comput.*, **30** (2018), 1–10. https://doi.org/10.1287/ijoc.2017.0759

27. D. Efrosinin, N. Stepanova, J. Sztrik, Algorithmic analysis of finite-source multi-server heterogeneous queueing systems, *Mathematics*, **9** (2021), 2624. https://doi.org/10.3390/math9202624

28. J. Roszik, J. Sztrik, Performance analysis of finite-source retrial queues with nonreliable heterogenous servers, *J. Math. Sci.*, **146** (2007), 6033–6038.

29. Q. M. He, H. Li, Y. Q. Zhao, Ergodicity of the $BMAP/PH/s/s+K$ retrial queue with PH-retrial time, *Queueing Syst.*, **35** (2000), 323–347. https://doi.org/10.1023/A:1019110631467

30. L. Breuer, A. Dudin, V. Klimenok, A retrial $BMAP/PH/N$ system, *Queueing Syst.*, **40** (2002), 433–457. https://doi.org/10.1023/A:1015041602946

31. L. Breuer, V. Klimenok, A. Birukov, A. Dudin, U. R. Krieger, Modeling the access to a wireless network at hot spots, *Eur. Transact. Telecommun.*, **16** (2005), 309–316. https://doi.org/10.1002/ett.1000

32. A. N. Dudin, R. Manzo, R. Piscopo, Single server retrial queue with group admission of customers, *Comput. Oper. Res.*, **61** (2015), 89–99. https://doi.org/10.1016/j.cor.2015.03.008

33. A. N. Dudin, S. A. Dudin, R. Manzo, L. Rarità, Analysis of multi-server priority queueing system with hysteresis strategy of server reservation and retrials, *Mathematics*, **10** (2022), 3747. https://doi.org/10.3390/math10203747

34. C. D'Apice, M. P. D'Arienzo, A. Dudin, R. Manzo, Admission control in priority queueing system with servers reservation and temporal blocking admission of low priority users, *IEEE Access*, **11** (2023), 44425–44443. https://doi.org/10.1109/ACCESS.2023.3273148

35. S. A. Dudin, O. S. Dudina, Call center operation model as a $MAP/PH/N/N-R$ system with impatient customers, *Probl. Inform. Transmiss.*, **47** (2011), 364–377. https://doi.org/10.1134/S0032946011040053

36. H. Baumann, W. Sandmann, Multi-server tandem queue with Markovian arrival process, phase-type service times, and finite buffers, *Eur. J. Oper. Res.*, **256** (2017), 187–195. https://doi.org/10.1016/j.ejor.2016.07.035

37. A. N. Dudin, S. A. Dudin, R. Manzo, L. Rarità, Queueing system with batch arrival of heterogeneous orders, flexible limited processor sharing and dynamical change of priorities, *AIMS Math.*, **9** (2024), 12144–12169. https://doi.org/10.3934/math.2024593

38. 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

39. C. Kim, A. Dudin, O. Dudina, S. Dudin, Tandem queueing system with infinite and finite intermediate buffers and generalized phase-type service time distribution, *Eur. J. Oper. Res.*, **235** (2014), 170–179. https://doi.org/10.1016/j.ejor.2013.12.012

40. A. Dudin, C. Kim, S. Dudin, O. Dudina, Priority retrial queueing model operating in random environment with varying number and reservation of servers, *Appl. Math. Comput.*, **269** (2015), 674–690. https://doi.org/10.1016/j.amc.2015.08.005

41. 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