



Article

# Self-Service System with Rating Dependent Arrivals

Alexander Dudin <sup>1,2,\*</sup> , Olga Dudina <sup>1</sup>, Sergei Dudin <sup>1</sup> and Yulia Gaidamaka <sup>2</sup> 

<sup>1</sup> Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus; dudina@bsu.by (O.D.); dudins@bsu.by (S.D.)

<sup>2</sup> Applied Mathematics and Communications Technology Institute, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St., 117198 Moscow, Russia; gaydamaka-yuv@rudn.ru

\* Correspondence: dudin-alexander@mail.ru

**Abstract:** A multi-server infinite buffer queueing system with additional servers (assistants) providing help to the main servers when they encounter problems is considered as the model of real-world systems with customers' self-service. Such systems are widely used in many areas of human activity. An arrival flow is assumed to be the novel essential generalization of the known Markov Arrival Process (MAP) to the case of the dynamic dependence of the parameters of the MAP on the rating of the system. The rating is the process defined at any moment by the quality of service of previously arrived customers. The possibilities of a customers immediate departure from the system at the entrance to the system and the buffer due to impatience are taken into account. The system is analyzed via the use of the results for multi-dimensional Markov chains with level-dependent behavior. The transparent stability condition is derived, as well as the expressions for the key performance indicators of the system in terms of the stationary probabilities of the Markov chain. Numerical results are provided.

**Keywords:** multi-server queueing model; rating; self-sufficient servers; self-checkout; assistants; multi-dimensional Markov chains



**Citation:** Dudin, A.; Dudina, O.; Dudin, S.; Gaidamaka, Y. Self-Service System with Rating Dependent Arrivals. *Mathematics* **2022**, *10*, 297. <https://doi.org/10.3390/math10030297>

Academic Editor: János Sztrik

Received: 11 December 2021

Accepted: 17 January 2022

Published: 19 January 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Queueing theory is very useful for modeling various real-world systems, contact centers, airports, banks, telecommunication, and retail networks, in particular. The queueing model considered in this paper has two main novel features: (i) the mechanism of customer's arrival is dependent on the current rating of the system and (ii) consideration of self-service of customers via so-called self-service devices (SSD) or self-checkouts. Both these features are inherent for many real systems, e.g., entertainment systems, contact centers, food services, and retail networks. Thus, they have to be carefully taken into account in system design and management aiming to guarantee the effective operation of a system. The effective operation suggests earning the maximal profit received by the system via customers service and a high degree of customer satisfaction.

The standard queueing models considered in the literature suggest that the arrival flow of customers to the system does not depend on the system state. The flow entering the service may depend on this state via the mechanism of customers admission depending on the visibility of a queue, as well as its length and (or) the number of busy servers. In this paper, we assume the dependence of the arrival process not on the queue length but the rating of the system.

The rating is now the well-known notion that reflects the customers' satisfaction and has an influence on the choice among the competitive service systems in which the customer can obtain service. Ratings of various service systems are now very popular and easily available, e.g., on the Internet. Checking ratings/reviews before making consumption decisions has become a ritual for many of today's customers, see [1]. The rating dynamically evaluates the current quality of customer service in this system. The rating of the system

may cause so-called word-of-mouth advertising and has quite a strong influence on the preferences of the customers and their arrival rate to any system. In turn, this has a high impact on the profit earned by this system, and the ratings have to be taken into account in the management of the system operation. Analysis of queueing systems with an arrival process depending on the rating of the system is of high practical importance.

The existing literature about the queues with servers rating is not extensive. In [1], a queueing model with two types of customers is considered. Sophisticated customers are well-informed of service-related information and make their joining-or-balking decisions strategically, whereas naive customers do not have such information and rely on online rating information to make such decisions. The problem of optimal pricing strategy is solved in [1]. The queueing model containing two competitive systems with customers' choice of the system to be joined via the comparison of the individual ratings of the systems was recently considered with the use of matrix analytic methods in [2].

In modern retail networks, hotels, banks, airports, etc., there is a robust trend for extending the use of self-service devices (*SSD*) or self-checkouts. *SSDs* have been defined by the new technological interfaces (e.g., the Quick Response (*QR*) codes, image and face recognition, radio frequency identification (*RFID*)) that allow customers to produce services without a service employee's involvement, see, e.g., [3]. The human operator (administrator, assistant, etc.) is involved in the service only upon request of a customer who asks for help in resolving certain problems that he/she met during a service or in the case of violation of the established rules by a customer. The use of *SSD* is becoming very wide nowadays, in particular, in many retail networks worldwide due to many reasons. The main reason is that it is profitable for both owners of services and customers.

The owners save money via non-payment of salaries to service employees and other operational costs. This creates better opportunities for successful competition with very popular now online shopping that is associated for many customers with the safest (from the perspective of health safety) and convenient way of shopping. Statistics show that one human operator (administrator, assistant, etc.) can easily control 6–10 *SSDs*. The *SSDs* take up less space than the regular cash registers, which allows optimizing the store space. With their help, it is possible to unload the cash register area and to increase its throughput. Among other things, *SSDs* encourage customers to make additional purchases. At one time, McDonald's found out during an experiment that visitors spend an average of 30 percent more on purchases when they are not worried that the person behind the cash register will evaluate their choice. The use of *SSDs* may allow the reduction of the actual and perceived waiting time that is strongly linked with customers satisfaction. In turn, this should imply higher loyalty of customers and the future profit of the owner.

The main profit gained by a customer consists of: (i) having a chance to avoid long waiting in the queue until the human server (cashier) will become available. Waiting generally is regarded as an undesirable activity that customers must undertake to complete the service. Waiting can lead to both emotional (anger, irritation, frustration, boredom, stress) and behavioral (e.g., abandonment or renegeing) responses, especially when it is costly and limits the person's ability to engage in more productive or rewarding ways to spend their time; (ii) getting more control over his/her shopping experience; (iii) obtaining a possibility of better distancing from other buyers what is very important in the current era of the COVID-19 pandemic. With the continuous improvement in technology and the promotion of self-service retail stores in the market, their numbers will increase. Furthermore, the scales of users and transactions will rapidly increase in the future. For more existing literature about the perspectives and attractiveness of the use of *SSDs*, see, e.g., [4–8].

In the early beginning, the spread of the use of *SSDs* was fully justified by the huge investment of the companies to enforce the use of the new perspective technologies. After they are already implemented in life, it is necessary to effectively manage the operation of each concrete service system. To this end, besides many administrative problems, a whole bunch of pure mathematical problems has to be resolved. One of these problems is a traditional problem in modeling service systems. Namely, given the actual or expected

characteristics of the arrival process, the distribution of service time of one customer, and the required values of the service level indicators, it is necessary to optimally choose the number of required servers (*SSDs*). Such indicators may be, e.g., the probability that the waiting time of an arbitrary customer will not exceed the given value with the fixed in advance value or the probability of customer abandonment.

It is known that, during the use of *SSD* for purchases, the customers can meet problems related, e.g., to the search of the necessary good on the shelves, readability of RFIDs, correct use of the scales, damage of the goods. To resolve the potentially arising problems, usually, the stores have some additional staff of administrators (assistants, helpers, etc.). They provide help to a customer if at least one of the assistants is not busy. Otherwise, the customer should wait until one of the assistants becomes available. Therefore, the problem of the optimal choice of the number of *SSDs* is supplemented by the problem of the optimal matching of the number of required assistants to the number of the *SSDs*. The redundant number of assistants implies higher, unjustified operational costs. The insufficient number of assistants causes long waiting times for help for a customer. That, in turn, implies longer total service time of this customer, longer waiting time of other customers, the higher probability of the abandonment and reneging by an arbitrary customer from the system, and the loss of potential profit that could be earned by service of customers. In this paper, we solve the problem of computation of the values of performance indicators under any fixed set of the numbers of *SSDs* and assistants. The problem is formulated and solved in borders of the matrix queueing theory. The usability of this result for the optimal choice of such a set is numerically illustrated.

Due to the practical importance of the effective use of *SSDs*, there are a lot of papers devoted to this topic. We mention only a few of them that operate with the notion of a customer waiting time. Analysis of the waiting time is one of the standard goals in queueing theory. In [6], the usefulness of queueing theory for the analysis of systems with *SSD* is noted. The question of the relation of the actual waiting time of a customer and the perceived waiting time, as well as their strong link with customer satisfaction, are discussed. Customer satisfaction is strongly associated with the loyalty of the customers, which is very important for service providers. Therefore, analysis of the ways to increase the loyalty of the customers strongly correlates with an analysis of the actual waiting time of a customer. Such a time is one of the key performance indicators of the majority of queueing systems. Thus, queueing analysis is an important part of solving the problem of the optimal design of the systems of *SSDs*. However, the analysis of many queueing systems is quite complicated. This explains why analysis of these systems is often implemented not via the use of the analytical and algorithmic methods of queueing theory but via the computer simulation methods. Namely, the method of computer simulation is used for the experimental study of the systems of *SSDs*. In [7], the correlation of the waiting time, customer experience, and satisfaction was discussed via the use of certain methods of sociology. The content of [8,9] is similar to [7], and the methods of sociology are also used.

In our paper, we provide an analysis of the queueing model with *SSDs* and assistants. The existing literature about the queues with service assistants is quite scarce. The recent paper [10] is devoted to the analysis of the set of *SSDs* described in terms of a tandem queueing model with a single-server first phase and a multi-server second phase. All distributions defining the system operation are exponential. The behavior of the system is described by a two-dimensional Markov chain that is the Quasi-Birth-and-Death-Process. This process is easily analyzed via the tools of the matrix-geometric method by M. Neuts, see [11].

The queueing model of the self-checkout (self-service) system considered in our paper assumes the existence of two multi-server sub-systems. Let us denote the number of servers in these two systems as  $N$  and  $M$ , correspondingly. The first sub-system defines the service process of customers by themselves. Any arriving customer that does not abandon the system (due to too long, in his/her opinion, queue) obtains service in this sub-system and successfully departs from the system if he/she does not encounter service problems.

If a problem occurs, the server from the second sub-system has to help in resolving this problem if they are available. If all servers of the second sub-system are busy, the customer that encountered the problem suspends their service until any server of the second sub-system becomes available. After the problem is resolved, the server in the first sub-system resumes the work while the corresponding assistant is released. A problem in service at the first sub-system can occur an arbitrary number of times. After service is completed, the customer departs from the system.

As follows from this brief description, our model does not belong to the class of tandem queues because the service of an arbitrary customer does not strictly consist of at most two sequential services at different queueing systems. This service may be the sequence of alternating services by servers from the first and the second sub-systems. Our model is more similar to the unreliable queue with repairmen. The first sub-system describes the service of customers, and the second sub-system describes the behavior of the pool of repairmen. However, the overwhelming majority of the papers devoted to this subject consider only the joint distribution of the number of non-broken servers and the number of busy repairmen. The duty of servers to provide service and a possible queue of customers are not taken into account; see the survey [12], paper [13], and references therein. In our model, namely, the characteristics of the customers' service quality are in the focus of the study.

Models similar to our queueing models (however, without the rating consideration) are considered in the following papers. In [14], the model with  $N$  servers and  $M = 1$  assistants (called in [14] as the main servers and the consultant) is considered. Arrivals are defined by quite a general Markov Arrival Process (*MAP*); for definition, properties, and related research, see [15–18]. Other distributions characterizing the system are assumed to be exponential. The system is comprehensively analyzed using the matrix analytic methods. Extensive illustrative numerical examples to bring out the qualitative nature of the model are presented. In [19], the model with  $N = 1$  servers and  $M = 1$  assistants is analyzed. All parameters of the system depend on the state of a finite state random environment. All involved distributions are assumed to be exponential. The system is analyzed using the matrix analytic methods. In [20], the model with  $N = 1$  servers and  $M = 1$  assistants is analyzed. The input buffer is finite, and the number of opportunities to ask for help by each server is restricted. Service and help times have so-called phase-type (*PH*) distributions, see [11]. The system is analyzed using the matrix analytic methods. The model considered in [21] assumes an arbitrary number of servers and assistants (called in this paper, specialist servers); the possibility of providing help by the assistant is only after the main service. Service cannot be continued after receiving help. Practically, this means the consideration of a tandem queueing model. The arrival process is the *MAP*, and the help times have *PH* distributions. The system is analyzed using the matrix analytic methods. A little bit similar to our model under quite general assumptions about the arrival process (the Batch Markov Arrival process) and service times (phase-time distribution) was recently considered in [22]. In that model, if the server does not succeed in finishing the service of a customer within a certain time, then the so-called backup server joins with the server for the service of this customer. When both servers serve a customer, the service speed increases. Another difference is that in [22], after obtaining help from the backup server, the server mandatorily finishes the service. In our model, we suggest that the server can obtain help from the assistant many times, and the server and the assistant do not cooperate in service. During obtaining the help, service is not provided.

The additional two features of the model considered in this paper are the following.

- The model suggests the visible queue. This means that the arriving customer can see the number of customers in the queue and can use this information to decide to join a queue or abandon (balk) it, e.g., queueing systems managed by ticket technology is widely used in service industries, as well as government offices, see [23]. Upon arriving at a ticket queue, each customer is issued a numbered ticket. The number currently being served is displayed. An arriving customer balks if the difference

between their ticket number and the displayed number exceeds their patience level. Analysis of a queueing model in [23] was implemented via the use of matrix analytic methods. Systems with a visible queue were also previously considered, e.g., in [24,25]. In [24], an empirical study of queue abandonment by the patients in an emergency department of a hospital is implemented by the methods of econometrics. In [25], the queueing/inventory model with a visible queue was analyzed via the use of matrix analytic methods.

- The model suggests the impatience of customers waiting in the buffer. Abandonment or balking means the refusal of a customer from joining the queue due to it being inappropriate for him/her in length. Another possible reason for customer loss is his/her impatience or renegeing. Initially, a customer joins the queue. However, if his/her waiting time exceeds some critical (deterministic or random) level, the customer departs from the system. The phenomenon of impatience is important and received quite a lot of attention in the literature; for more references, see, e.g., [26–28]. In [27], in particular, the problem of wasting time by the server because the arriving customer decides to join a queue and receives a ticket but then departs from the system without canceling the ticket is considered. In [29], the authors analyzed the model in which customers' patience is exponentially distributed, and the system's waiting capacity is unlimited. Such a model is both rich and analyzable enough to provide information that is practically important for call center managers. The distinguishing feature of the paper [30] is that service time distributions in the considered multi-server queueing model are generally distributed. A simple and insightful solution is presented for the loss probability. The solution offered in [30] is exact for exponential services and is an excellent heuristic for general service times. In [31], a single server variant of the model from [30] is under study. In [32], the customer's loss probability in  $M/M/c$  queue with impatient customers is expressed in a simple formula involving the waiting time probabilities in the  $M/M/c$  queue with patient customers. In that paper, a probabilistic derivation of this formula is given, and possible use of this general formula in the  $M/M/c$  retrial queue with impatient customers is outlined. In [33–39], the retrial models with impatient customers are also considered. The model considered in [40] assumes that the multi-server queue operates under the influence of the random environment, and the impatience rate depends on the current state of the random environment. In [41–43], multi-server queues with the *MAP* or marked *MAP* arrival flows and impatient customers as the models of call centers were analyzed via the matrix analytic methods.

The remainder of the text of this paper consists of the following. In Section 2, the mathematical model is completely described. In Section 3, the process describing the dynamics of the system is defined as the continuous-time multi-dimensional Markov chain with level-dependent transitions. The generator of this chain is given. Section 4 contains the ergodicity and non-ergodicity conditions for this Markov chain in the transparent form. Section 5 briefly touches on the question of computation of the stationary distribution of the Markov chain. In Section 6, expressions for computation of the key performance indicators of the system in terms of the computed stationary probabilities of the system states are presented. Section 7 contains the results of the numerical experiment, the aim of which is to give insights into the shape of dependencies of the performance indicators on the number of servers and assistants. It is worth noting that the numerical realization of the elaborated paper algorithm results, are implemented not for a toy example with a small number of servers and assistants but for the system with realistic numbers of *SSDs* and assistants. Section 8 concludes the paper.

## 2. Mathematical Model

We consider a multi-server queueing system with an infinite buffer having two types of servers. The structure of the system is presented in Figure 1.



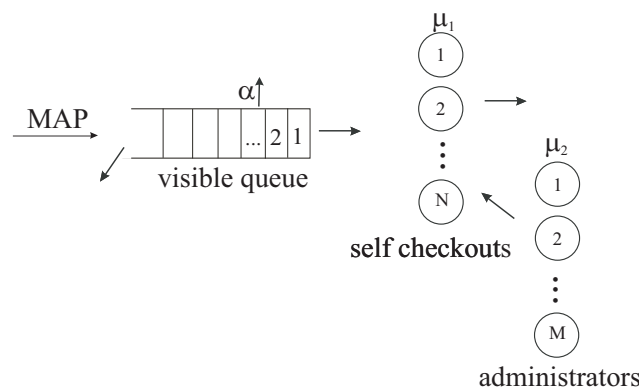


Figure 1. Structure of the system.

The servers of the first type correspond to self-checkouts (self-service devices). The total number of such servers is equal to  $N$ . The servers of the second type correspond to administrators (assistants). The total number of assistants is equal to  $M$ . If a customer is accepted for service, he/she firstly occupies one first-type server for service during an exponentially distributed time with the parameter  $\mu_1$ . After this time expires, with a probability of  $1 - p$ , the customer successfully finishes service and departs from the system. With the complementary probability, the customer meets a problem with service and requires some kind of help from an administrator. If at this moment, there is a free administrator, this administrator provides help to the customer to resolve the problem during an exponentially distributed with the parameter  $\mu_2$  time. After that, the customer continues his/her service during an exponentially distributed time with the parameter  $\mu_1$ . The number of moments during service when a customer asks the help of an administrator (not necessarily the same as in the previous moments) is unlimited. If the help of an administrator is required while there is no free administrator, the customer suspends their service and waits until any administrator will become available.

The process of customers' arrival to the system is the generalization of the Markovian arrival process (MAP), see, e.g., [15–18], described below. Arrivals in the classical MAP are governed by the underlying Markov chain  $v_t, t \geq 0$ , with the state space  $\{1, 2, \dots, W\}$ . The generator  $D$  of this chain is represented in the additive form  $D = D_0 + D_1$ , where the components of the matrix  $D_1$  define intensities of transitions of the chain that are accompanied by the customer's arrival. The non-diagonal components of the matrix  $D_0$  define the intensities of transitions of the chain that are not accompanied by the customer's arrival. The diagonal components are negative. The moduli of these components define the rates of the exit from the corresponding state of the Markov chain.

In contrast to a classical MAP, we assume that the transition intensities of the underlying process  $v_t$  additionally depend on the parameter  $r, r = \overline{1, R}$ , which defines the so-called current rating of the system. If the rating of the system is  $r$ , then the MAP is characterized by the square matrices  $D_0^{(r)}$  and  $D_1^{(r)}$  of size  $W$ . The average arrival rate of the MAP when the rating of the system is  $r$  is denoted as  $\lambda_r$ , which can be found as  $\lambda_r = \theta^{(r)} D_1^{(r)} \mathbf{e}, r = \overline{1, R}$ , where  $\theta^{(r)}$  is an invariant vector of the MAP defined by the matrices  $D_0^{(r)}$  and  $D_1^{(r)}$ , and  $\mathbf{e} = (1, 1, \dots, 1)^T$ . We do not specify the concrete form of the matrices  $D_0^{(r)}$  and  $D_1^{(r)}$ . We only suggest that the increase in the rating cannot imply the decrease in the average arrival rate, i.e., we require that the following inequalities hold true:

$$\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_R.$$

The rating of the system can be dynamically changed during the system operation. We assume that if a customer is admitted to the system without waiting in the queue, its current rating  $r, r = \overline{1, R - 1}$ , immediately increases by one with the fixed small probability  $r^+$ . For example, if we fix  $r^+ = 0.001$ , the service of 1000 customers without waiting in the

queue leads, on average, to the increase in the rating by one. If a customer abandons the system without service, the current rating  $r$ ,  $r = \overline{2}, \overline{R}$ , decreases by one with the probability  $r^-$ . For example, if we fix  $r^- = 0.01$ , the loss of 100 customers leads, on average, to the decrease in the current rating by one. At the moment of the change of the rating, transition intensities of the underlying process  $\nu_t$  immediately adjust their values to the new rating.

The queue is assumed to be visible. This means that an arriving customer can observe the number of customers in the queue and leave the system without service if he/she considers this number inappropriate. We assume that if during an arbitrary customer arrival epoch there is no free server and the number of customers in the buffer is  $i$ , the customer permanently leaves the system with the probability  $q_i$ ,  $i \geq 0$ . With the complementary probability, the customer joins the queue. We suggest that the limit  $q = \lim_{i \rightarrow \infty} q_i$  exists,  $0 < q \leq 1$ . Additionally, we assume that customers can be impatient and leave the buffer and depart from the system, independently of each other, after an exponentially distributed with the parameter  $\alpha$ ,  $\alpha \geq 0$ , amount of time. In the case of  $\alpha = 0$ , the customers are patient.

Now, let us analyze the described queueing model.

### 3. Process of the System States

The behavior of the system under study can be described by the regular irreducible continuous-time Markov chain

$$\zeta_t = \{i_t, n_t, r_t, \nu_t\}, t \geq 0,$$

where, during the epoch  $t$ ,

- $i_t$  is the number of customers in the system,  $i_t \geq 0$ ;
- $n_t$  is the number of blocked (waiting for help from an assistant) servers,  $n_t = \overline{0, \min\{i_t, N\}}$ ;
- $r_t$  is the current rating of the system,  $r_t = \overline{1, \overline{R}}$ ;
- $\nu_t$  is the state of the underlying process of the MAP,  $\nu_t = \overline{1, \overline{W}}$ .

Here and further, the notation  $n = \overline{0, N}$  means that the parameter  $n$  admits values from the set  $\{0, \dots, N\}$ .

To formally define the continuous-time Markov chain  $\zeta_t$ , it is necessary to write down, for any pair of the states  $(i, n, r, \nu)$  and  $(i', n', r', \nu')$ , the intensity of the transitions between these states.

To avoid bulky denotations, following the standard methodology of investigation of multi-dimensional Markov chains having one denumerable component, we enumerate the states of the Markov chain  $\zeta_t = \{i_t, n_t, r_t, \nu_t\}$  in the direct lexicographic order of the components  $\{n_t, r_t, \nu_t\}$  and combine the set of the states with the value  $i$  of the component  $i_t$  into the so-called level  $i$ ,  $i \geq 0$ .

Let  $Q_{i,j}$  be the matrix constituted by the transition intensities from level  $i$  to level  $j$  and let  $Q$  be the block matrix constituted by the blocks  $Q_{i,j}$ ,  $i \geq 0$ ,  $j \geq 0$ . It is clear that the matrix  $Q$  is the infinitesimal generator of the Markov chain  $\zeta_t$ ,  $t \geq 0$ .

**Theorem 1.** *The generator  $Q$  of the Markov chain  $\zeta_t$ ,  $t \geq 0$ , has the following block three-diagonal structure*

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

The non-zero blocks are defined as follows:

$$Q_{0,0} = \hat{D}_0,$$

$$Q_{i,i} = I_{i+1} \otimes \hat{D}_0 + (-\mu_2 C_i - \mu_1 \tilde{C}_i + \mu_2 C_i E_i^- + p\mu_1 \tilde{C}_i E_i^+) \otimes I_{RW}, 0 < i < N,$$

$$\begin{aligned}
 Q_{i,i} &= I_{N+1} \otimes \hat{D}_0 + q_{i-N} I_{N+1} \otimes \hat{D}_1 ((r_- \mathcal{R}^- + (1 - r_-) I_R) \otimes I_W) + (-\mu_2 C_N - \mu_1 \tilde{C}_N + \\
 &\quad + \mu_2 C_N E_N^- + p \mu_1 \tilde{C}_N E_N^+) \otimes I_{RW} - (i - N) \alpha I_{(N+1)RW}, \quad i \geq N, \\
 Q_{i,i+1} &= \tilde{E}_i \otimes \hat{D}_1 ((r_+ \mathcal{R}^+ + (1 - r_+) I_{R+1}) \otimes I_W), \quad 0 < i < N, \\
 Q_{i,i+1} &= (1 - q_{i-N}) I_{N+1} \otimes \hat{D}_1, \quad i \geq N, \\
 Q_{i,i-1} &= \mu_1 (1 - p) \tilde{C}_i \hat{E}_i \otimes I_{RW}, \quad 0 < i \leq N, \\
 Q_{i,i-1} &= (i - N) \alpha I_{(N+1)} \otimes ((r_- \mathcal{R}^- + (1 - r_-) I_R) \otimes I_W) + \mu_1 (1 - p) \tilde{C}_N \otimes I_{RW}, \quad i > N,
 \end{aligned}$$

where

$\otimes$  indicates the symbol of the Kronecker product matrices, see [44];

$\hat{D}_0 = \text{diag}\{D_0^{(1)}, D_0^{(2)}, \dots, D_0^{(R)}\}$ , where  $\text{diag}\{\dots\}$  denotes the diagonal matrix with the diagonal entries listed in the brackets;

$\hat{D}_1 = \text{diag}\{D_1^{(1)}, D_1^{(2)}, \dots, D_1^{(R)}\}$ ;

$C_i = \text{diag}\{0, 1, \dots, \min\{i - 1, M\}, \min\{i, M\}\}$ ,  $i = \overline{1, N}$ ;

$\tilde{C}_i = \text{diag}\{i, i - 1, \dots, 0\}$ ,  $i = \overline{1, N}$ ;

$I_K$  is the identity matrix having a size indicated in the suffix (if the size of the matrix is clear from the context, it can be omitted);

$O_K$  is a zero matrix having a size indicated in the suffix (if the size of the matrix is clear from the context, it can be omitted);

$E_i^-$  is a square matrix of size  $i + 1$  with all zero entries, except the entries  $(E_i^-)_{l,l-1}$ ,  $l = \overline{2, i + 1}$ ,  $i = \overline{1, N}$ , which are equal to 1;

$E_i^+$  is a square matrix of size  $i + 1$  with all zero entries, except the entries  $(E_i^+)_{l,l+1}$ ,  $l = \overline{1, i}$ ,  $i = \overline{1, N}$ , which are equal to 1;

$\tilde{E}_i$  is a matrix of size  $(i + 1) \times (i + 2)$  with all zero entries, except the entries  $(\tilde{E}_i)_{l,l}$ ,  $l = \overline{1, i + 1}$ ,  $i = \overline{1, N}$ , which are equal to 1;

$\hat{E}_i$  is a matrix of size  $(i + 1) \times i$  with all zero entries, except the entries  $(\hat{E}_i)_{l,l}$ ,  $l = \overline{1, i}$ ,  $i = \overline{1, N}$ , which are equal to 1;

$\mathcal{R}^+$  is a matrix of size  $R \times R$  with all zero entries, except the entries  $(\mathcal{R}^+)_{l,l+1}$ ,  $l = \overline{1, R - 1}$ , and  $(\mathcal{R}^+)_{R,R}$ , which are equal to 1;

$\mathcal{R}^-$  is a matrix of size  $R \times R$  with all zero entries, except the entries  $(\mathcal{R}^-)_{l,l-1}$ ,  $l = \overline{2, R}$ , and  $(\mathcal{R}^-)_{1,1}$ , which are equal to 1.

**Proof.** The proof of Theorem 1 is implemented via careful analysis of all possible transitions of the Markov chain  $\zeta_t$ ,  $t \geq 0$ , and further combining the intensities of these transitions into the blocks of the generator.

The generator  $Q$  has all negative diagonal entries and non-negative non-diagonal entries. The diagonal entries of the generator  $Q$  define, up to the sign, the total intensity of leaving the corresponding state of the Markov chain  $\zeta_t$ ,  $t \geq 0$ . In the case  $i = 0$ , the Markov chain can leave the current state only if the underlying process of the arrival process makes a transition. The intensities of such transitions are defined as the modules of the diagonal entries of the matrix  $\hat{D}_0$ .

In the case when the system is not idle, but the buffer is empty, the Markov chain  $\zeta_t$ ,  $t \geq 0$ , can also change its state due to the finish of the service by a server or finish of help provided by an assistant. The intensities of such transitions are defined as the diagonal entries of the matrix  $(\mu_2 C_i + \mu_1 \tilde{C}_i) \otimes I_{RW}$ ,  $i = \overline{1, N}$ .

If the number of customers in the buffer is greater than zero, the Markov chain  $\zeta_t$ ,  $t \geq 0$ , can also change its state due to the departure of some customers from the buffer due to impatience. The intensities of such transitions are defined as the diagonal entries of the matrix  $(i - N) \alpha I_{(N+1)RW}$ ,  $i > N$ .

The non-diagonal entries of the matrices  $Q_{i,j}$ ,  $i \geq 0$ , define the intensities of transitions that do not lead to the change of the number of customers in the system  $i$ . Such transitions are the following:



(1) The underlying process of the arrival process transition, which does not imply the acceptance of a new customer to the system (the customer is not generated or lost). The intensities of such transitions are given as the non-diagonal entries of the matrix  $I_{\min\{N,i\}+1} \otimes \hat{D}_0$  and the entries of the matrix  $q_{i-N} I_{N+1} \otimes \hat{D}_1((r_- \mathcal{R}^- + (1 - r_-) I_R) \otimes I_W)$  for  $i \geq N$ . Note, that here the matrix  $(r_- \mathcal{R}^- + (1 - r_-) I_R)$  defines the possible change of the system rating due to the loss of a customer;

(2) The number of blocked servers is increased by one. The intensities of such transitions are given as the entries of the matrix  $(p\mu_1 \tilde{C}_{\min\{N,i\}} E_{\min\{N,i\}}^+) \otimes I_{RW}$ ;

(3) The number of blocked servers is decreased by one. The intensities of such transitions are given as the entries of the matrix  $(\mu_2 C_{\min\{N,i\}} E_{\min\{N,i\}}^-) \otimes I_{RW}$ .

The entries of the matrices  $Q_{i,i+1}$ ,  $i \geq 0$  define the intensities of transitions that lead to the increase in the number of customers  $i$  in the system by one. This can happen only in the case when the underlying arrival process makes a transition with the generation of a customer, and this customer is admitted to the system. The intensities of such transitions are given as the entries of the matrices  $\tilde{E}_i \otimes \hat{D}_1((r_+ \mathcal{R}^+ + (1 - r_+) I_{R+1}) \otimes I_W)$ , in the case  $i = \overline{1, N-1}$ , and the entries of the matrices  $(1 - q_{i-N}) I_{N+1} \otimes \hat{D}_1$ , if  $i \geq N$ .

The entries of the matrices  $Q_{i,i-1}$ ,  $i \geq 1$ , define the intensities of transitions that lead to the decrease in the number of customers  $i$  in the system by one. This can happen if a customer leaves the system successfully serviced (the intensities of such transitions are given as the entries of the matrices  $\mu_1(1 - p) \tilde{C}_{\min\{N,i\}} \hat{E}_{\min\{N,i\}} \otimes I_{RW}$ ) and if a customer leaves the non-empty buffer due to impatience (the intensities of such transitions are given as the entries of the matrices  $(i - N)\alpha I_{(N+1)} \otimes ((r_- \mathcal{R}^- + (1 - r_-) I_R) \otimes I_W)$ ).

The blocks  $Q_{i,j}$ ,  $i, j \geq 0$ ,  $|i - j| > 1$ , of the generator  $Q$  are zero matrices because the customers arrive and leave the system only one-by-one.  $\square$

#### 4. Ergodicity Condition

An important step of analysis of any Markov chain with an infinite state space is establishing conditions for the ergodicity and non-ergodicity of this Markov chain (the stability condition). The ergodicity and non-ergodicity conditions for the Markov chain under study  $\xi_t$  are given by the following Theorem.

**Theorem 2.** (a) If the impatience rate  $\alpha$  is positive, the Markov chain  $\xi_t$  is ergodic under all finite values of other parameters of the considered queueing system;

(b) If the limiting probability  $q$  is equal to 1, the Markov chain  $\xi_t$  is ergodic under all finite values of other parameters of the considered queueing system;

(c) If the customers staying in the buffer are patient, i.e.,  $\alpha = 0$ , the Markov chain  $\xi_t$  is ergodic if the following inequality is fulfilled:

$$\lambda_1(1 - q) < \sum_{n=0}^N \gamma_n(N - n)\mu_1, \tag{1}$$

where  $\gamma_n$  are the probabilities that at an arbitrary moment when the system is overloaded, the number of assistants providing help to the servers is equal to  $n$ ,  $n = \overline{0, N}$ .

These probabilities are computed by the formula:

$$\gamma_n = \left( 1 + \sum_{j=1}^N \prod_{l=1}^j \frac{p(N - l + 1)\mu_1}{l\mu_2} \right)^{-1} \prod_{l=1}^n \frac{p(N - l + 1)\mu_1}{l\mu_2}, \quad n = \overline{0, N}; \tag{2}$$

(d) If the customers staying in the buffer are patient, i.e.,  $\alpha = 0$ , the Markov chain  $\xi_t$  is non-ergodic if the following inequality is fulfilled:

$$\lambda_1(1 - q) > \sum_{n=0}^N \gamma_n(N - n)\mu_1. \tag{3}$$

**Proof.** Let the generator of a Markov chain be the upper-Hessenbergian matrix, i.e., it has zero blocks below the first sub-diagonal and other blocks  $Q_{i,i+k-1}$ ,  $k \geq 0$ . Let the matrix  $T_i$  be the diagonal matrix, the diagonal entries of which also coincide with the modules of the diagonal entries of the matrix  $Q_{i,i}$ ,  $i \geq 0$ .

If the following limits exist:

$$Y^{(k)} = \lim_{i \rightarrow \infty} \mathcal{R}_i^{-1} Q_{i,i+k-1}, \quad k = 0, 2, 3, \dots, \quad Y^{(1)} = \lim_{i \rightarrow \infty} \mathcal{R}_i^{-1} Q_{i,i} + I,$$

and the matrix  $\sum_{k=0}^{\infty} Y^{(k)}$  is stochastic; then the Markov chain belongs to the class of Asymptotically Quasi-Toeplitz Markov chains (AQTCM), see [45].

The generator  $Q$  of the Markov chain describing the queueing system under study defined by Theorem 1 has the particular, with respect to the upper-Hessenbergian matrix, three-block diagonal structure.

If  $\alpha > 0$ , then it can be verified that for this Markov chain the matrices  $Y_k$ ,  $k = 0, 1, 2$  exist and are defined by:  $Y^{(0)} = I$ ,  $Y^{(k)} = O$ ,  $k = 1, 2$ .

If  $\alpha = 0$ , then it can be verified that the matrices  $Y^{(k)}$  exist and are defined by

$$\begin{aligned} Y^{(0)} &= T^{-1} \tilde{Q}^-, \\ Y^{(1)} &= T^{-1} \tilde{Q}^0 + I, \\ Y^{(2)} &= T^{-1} \tilde{Q}^+, \end{aligned}$$

where

$$\begin{aligned} \tilde{Q}^- &= \mu_1(1 - p) \tilde{C}_N \otimes I_{RW}, \\ \tilde{Q}^0 &= I_{N+1} \otimes \hat{D}_0 + q I_{N+1} \otimes \hat{D}_1 ((r_- \mathcal{R}^- + (1 - r_-) I_R) \otimes I_W) + \\ &\quad (-\mu_2 C_N - \mu_1 \tilde{C}_N + \mu_2 C_N E_N^- + p \mu_1 \tilde{C}_N E_N^+) \otimes I_{RW}, \\ \tilde{Q}^+ &= (1 - q) I_{N+1} \otimes \hat{D}_1, \end{aligned}$$

and  $T$  is the diagonal matrix the diagonal entries of which coincide with the modules of the diagonal entries of the matrix  $\tilde{Q}^0$ .

Therefore, in both cases,  $\alpha > 0$  and  $\alpha = 0$ , the limits  $Y^{(k)}$ ,  $k = 0, 1, 2$ , exist and it is easy to check that their sum is the stochastic matrix. This implies that the considered Markov chain belongs to the class of AQTCM. The sufficient condition for the ergodicity of AQTCM, see [45], rewritten for the three-block diagonal generator is the fulfillment of the inequality:

$$\psi Y^{(0)} \mathbf{e} > \psi Y^{(2)} \mathbf{e} \tag{4}$$

where the vector  $\psi$  is the unique solution of equations

$$\psi (Y^{(0)} + Y^{(1)} + Y^{(2)}) = \psi, \quad \psi \mathbf{e} = 1. \tag{5}$$

The sufficient condition for the non-ergodicity is the fulfillment of the inequality

$$\psi Y^{(0)} \mathbf{e} < \psi Y^{(2)} \mathbf{e}.$$

Because for  $\alpha > 0$  we have  $Y^{(0)} = I$ ,  $Y^{(k)} = O$ ,  $k = 1, 2$ , inequality (4) takes a trivial form:  $1 > 0$ . Thus, the chain is ergodic for all finite values of other parameters of the considered queueing system. Statement (a) of the theorem is proven.

Consider now the case  $\alpha = 0$ . It can be verified that in this case system (4) and inequality (5) are equivalent to the system

$$\varphi (\tilde{Q}^- + \tilde{Q}^0 + \tilde{Q}^+) = \mathbf{0}, \quad \varphi \mathbf{e} = 1 \tag{6}$$

and the inequality

$$\varphi \tilde{Q}^- \mathbf{e} > \varphi \tilde{Q}^+ \mathbf{e}. \tag{7}$$

It can be verified that

$$\tilde{Q}^- + \tilde{Q}^0 + \tilde{Q}^+ = I_{N+1} \otimes ((1 - q)\hat{D}_1 + \hat{D}_0 + q\hat{D}_1((r_- \mathcal{R}^- + (1 - r_-)I_R) \otimes I_W)) + H$$

where

$$H = \begin{pmatrix} -pN\mu_1 & pN\mu_1 & O & \dots & O & O \\ \mu_2 & -\mu_2 - p(N-1)\mu_1 & p(N-1)\mu_1 & \dots & O & O \\ O & 2\mu_2 & -2\mu_2 - p(N-2)\mu_1 & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & N\mu_2 & -N\mu_2 \end{pmatrix} \otimes I_{RW}.$$

Using the so-called mixed product rule for Kronecker product of matrices, see [44], it is possible to show that a solution of (6) has the representation

$$\varphi = \Delta_1 \otimes \Delta_2 \tag{8}$$

where the row vector  $\Delta_1$  is a solution to the system

$$\Delta_1 \left[ I_{N+1} \otimes ((1 - q)\hat{D}_1 + \hat{D}_0 + q\hat{D}_1((r_- \mathcal{R}^- + (1 - r_-)I_R) \otimes I_W)) \right] = \mathbf{0}, \tag{9}$$

$$\Delta_1 \mathbf{e} = 1,$$

and the row vector  $\Delta_2$  is a solution to the system

$$\Delta_2 H = \mathbf{0}, \Delta_2 \mathbf{e} = 1. \tag{10}$$

By the direct substitution into (9), it is possible to check that vector  $\Delta_1$  defined by

$$\Delta_1 = (\theta^{(1)}, \mathbf{0}, \dots, \mathbf{0}) \tag{11}$$

is the solution of equation (9).

By the direct substitution into (10), it is possible to check that vector  $\Delta_2$  is defined by

$$\Delta_2 = (\gamma_0, \dots, \gamma_N) \tag{12}$$

where probabilities  $\gamma_n, n = \overline{0, N}$ , are given by Formula (2). Taking into account (8), (11), and (12) in (4), we obtain Formula (1).

It is clear that if  $q = 1$ , inequality (1) is always true. This proves statement (b) of the theorem. Statement (c) is also proven. The proof of statement (d) is easily made analogously to the proof of (b).  $\square$

**Remark 1.** Inequality (1) is intuitively transparent. Usually, the ergodicity condition is equivalent to the requirement that in the situation when the system is very overloaded, the rate of customers admission is less than the rate of customers service. The left-hand side of (1) is the rate of customers' admission to the system. As it is seen from (1), this rate here depends only on the arrival rate  $\lambda_1$  and does not depend on the rates  $\lambda_r, r = \overline{2, R}$ . This stems from the fact that the rating of the system, when it is very overloaded, is equal to 1 due to the abandonment (balking) of many customers. The right-hand side of (1) is the rate of customers' departure from the service when the system is overloaded. The values  $\gamma_n, n = \overline{0, N}$ , are the probabilities that  $n$  assistants provide help to the servers and these servers do not serve customers. Correspondingly, the number of servers providing service with the rate  $\mu_1$  is equal to  $N - n$ . It follows from the formula of total probability that the

right-hand side of (1) indeed is the rate of customers' departure from the service when the system is overloaded.

### 5. Computation of the Stationary Distribution of the Markov Chain

If the ergodicity condition is fulfilled, the stationary probabilities of the Markov chain  $\zeta_t, t \geq 0$ ,

$$\pi(i, n, r, v) = \lim_{t \rightarrow \infty} P\{i_t = i, n_t = n, r_t = r, v_t = v\},$$

$$i \geq 0, n = \overline{0, \min\{i, N\}}, r = \overline{1, R}, v = \overline{1, W};$$

exist.

We form the row vectors  $\pi_i, i \geq 0$ , of these stationary probabilities enumerated in the lexicographic order of the components  $(n, r, v), n = \overline{0, \min\{i, N\}}, r = \overline{1, R}, v = \overline{1, W}$ .

It is a well-known fact that these stationary probabilities can be found as the solution of the following system:

$$(\pi_0, \pi_1, \dots, \pi_N, \dots)Q = \mathbf{0},$$

$$(\pi_0, \pi_1, \dots, \pi_N, \dots)\mathbf{e} = 1.$$

Because the Markov chain  $\zeta_t, t \geq 0$ , is a level-dependent quasi-birth-and-death process having one countable component, this system cannot be solved using the standard matrix analytic methods. To solve this system, we recommend using the algorithms from papers [46,47].

### 6. Performance Indicators

The probability  $p_r, r = \overline{1, R}$ , that the value of the rating of the system at an arbitrary epoch is equal to  $r$  can be found as

$$p_r = \sum_{i=0}^{\infty} \sum_{n=0}^{\min\{i, N\}} \pi(i, n, r)\mathbf{e}.$$

The average rating  $\bar{R}$  of the system can be found as

$$\bar{R} = \sum_{r=1}^R r p_r.$$

The average customer's arrival rate  $\lambda$  can be found as

$$\lambda = \sum_{r=1}^R p_r \lambda_r.$$

The average output rate  $\lambda_{out}$  of successfully serviced customers can be found as

$$\lambda_{out} = \sum_{i=1}^{\infty} \sum_{n=0}^{\min\{i, N\}-1} (\min\{i, N\} - n)(1 - p)\mu_1 \pi(i, n)\mathbf{e}.$$

The average number  $N_{cust}$  of customers in the system is calculated as

$$N_{cust} = \sum_{i=1}^{\infty} i \pi_i \mathbf{e}.$$

The average number  $N_{buf}$  of customers in the buffer is calculated as

$$N_{buf} = \sum_{i=N+1}^{\infty} (i - N) \pi_i \mathbf{e}.$$

The average number  $N_{serv-1}$  of occupied servers is calculated as

$$N_{serv-1} = \sum_{i=1}^{\infty} \min\{i, N\} \pi_i \mathbf{e}.$$

The average number  $N_{blocked}$  of blocked servers is calculated as

$$N_{blocked} = \sum_{i=1}^{\infty} \sum_{n=1}^{\min\{i, N\}} n \pi(i, n) \mathbf{e}.$$

The average number  $N_{blocked-1}$  of blocked servers that are currently obtaining the help of an assistant is calculated as

$$N_{blocked-1} = \sum_{i=1}^{\infty} \sum_{n=1}^{\min\{i, N\}} \min\{n, M\} \pi(i, n) \mathbf{e}.$$

The average number  $N_{blocked-2}$  of blocked servers that are waiting until any assistant will become available is calculated as

$$N_{blocked-2} = \sum_{i=M+1}^{\infty} \sum_{n=M+1}^{\min\{i, N\}} (n - M) \pi(i, n) \mathbf{e} = N_{blocked} - N_{blocked-1}.$$

The average number  $N_{serv-2}$  of busy assistants is equal to  $N_{blocked-1}$ .

The loss probability  $P_{ent}$  of an arbitrary customer at the entrance to the system due to the unwillingness to wait in a long queue can be found as

$$P_{ent} = \frac{1}{\lambda} \sum_{i=N}^{\infty} \sum_{n=0}^N \sum_{r=1}^R q_{i-N} \pi(i, n, r) D_1^{(r)} \mathbf{e}.$$

The loss probability  $P_{imp}$  of an arbitrary customer due to impatience can be found as

$$P_{imp} = \frac{\alpha N_{buf}}{\lambda}.$$

The loss probability  $P_{loss}$  of an arbitrary customer can be found as

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

### 7. Numerical Example

The purpose of this example is to illustrate the dependencies of the main performance measures of the system on the number  $N$  of servers and the number  $M$  of assistants. Let us assume that the system can have up to 50 servers and up to 10 assistants. Thus, below we vary the parameter  $N$  over the interval  $(1, 50)$  and the parameter  $M$  over the interval  $(1, 10)$  with the same step 1.

We assume the following values of the parameters of the system:

- The service intensity of an arbitrary customer by a server is  $\mu_1 = 0.5$ .
- The probability  $p$  that a customer meets a problem with service and requires the help of an assistant is  $p = 0.25$ .
- The rate of help provided by an assistant is  $\mu_2 = 1.5$ .
- The parameter  $R$  that defines the maximum possible rating of the system is  $R = 10$ .
- The probability of rating increasing is  $r^+ = 0.001$ .
- The probability of rating decreasing is  $r^- = 0.005$ .
- The parameter  $\alpha$  describing the impatience rate of customers is equal to 0.06.

- The probabilities  $q_i = q_{i,N}$  that a customer will leave the system having  $N$  servers during the arrival epoch when the number of customers in the buffer is  $i$ , are defined as:

$$q_{i,N} = \begin{cases} \frac{i}{i+100N}, & \text{if } 0 \leq i \leq N; \\ \frac{i}{i+40N}, & \text{if } N < i \leq \max\{10, 2N\}; \\ \frac{i}{i+10N}, & \text{if } \max\{10, 2N\} < i \leq \max\{20, 5N\}; \\ \frac{i}{i+N}, & \text{if } \max\{20, 5N\} < i \leq \max\{100, 10N\}; \\ \frac{i}{i+0.1N}, & \text{otherwise.} \end{cases}$$

- The MAP arrival flow of customers when the system has the rating  $r$  is defined by the matrices  $D_0^{(r)} = rD_0^{base}$  and  $D_1^{(r)} = rD_1^{base}$ , where the matrices  $D_0^{base}$  and  $D_1^{base}$  are defined by

$$D_0^{base} = \begin{pmatrix} -2.5 & 0.02 \\ 0.001 & -0.8 \end{pmatrix}; D_1^{base} = \begin{pmatrix} 2.46 & 0.02 \\ 0.001 & 0.798 \end{pmatrix}.$$

The base arrival flow has the average intensity  $\lambda = 0.879048$ , the coefficient of correlation of successive inter-arrival times  $c_{cor} = 0.0557495$  and the coefficient of variation of inter-arrival times  $c_{var} = 1.12815$ .

The dependence of the average rating of the system  $\bar{R}$  on the parameters  $N$  and  $M$  is presented in Figure 2.

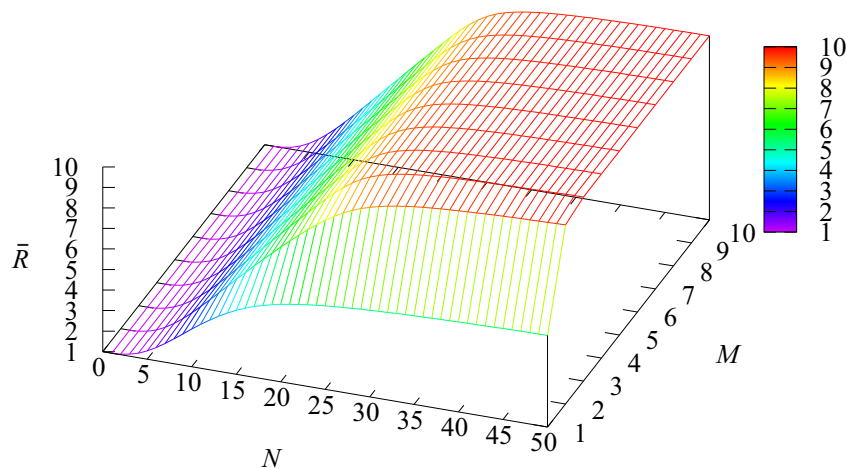


Figure 2. Dependence of the average rating of the system  $\bar{R}$  on  $N$  and  $M$ .

As is seen in Figure 2, the average rating increases with the increase in the number of servers and assistants. Here, we do not consider the situations when  $M \geq N$ . It is evident that in such a situation, adding a new assistant does not improve the system performance, and the rating is not changed. In other cases, adding new servers and (or) assistants leads to a better quality of customer service. As the result, the average rating increases.

Figure 3 illustrates the dependence of the average customer’s arrival rate  $\lambda$  on the parameters  $N$  and  $M$ .

The average arrival rate of customers also increases with the increase in the number of servers and (or) assistants. This is explained by the evident fact that a higher rating of the system leads to a higher arrival rate.

The dependence of the average number  $N_{buf}$  of customers in the buffer on the parameters  $N$  and  $M$  is presented in Figure 4.

The dependence of the average number  $N_{buf}$  of customers in the buffer on the parameters  $N$  and  $M$  is complicated and hardly predictable intuitively. On the one hand, as for a classical system, the increase in the number of servers leads to a decrease in the waiting time. However, for the considered system, the increase in the number of servers leads to an increase in the arrival rate that may imply an increase in the waiting time. As one can see



from the figure, sometimes the first factor prevails over the second one, and the average number  $N_{buf}$  of customers in the buffer decreases with the growth of  $N$  and  $M$ ; sometimes the situation changes oppositely.

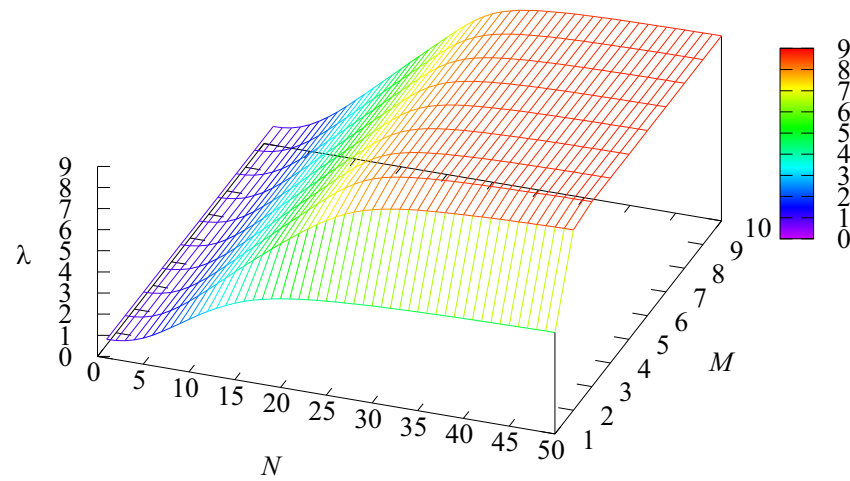


Figure 3. Dependence of the average customer’s arrival rate  $\lambda$  on  $N$  and  $M$ .

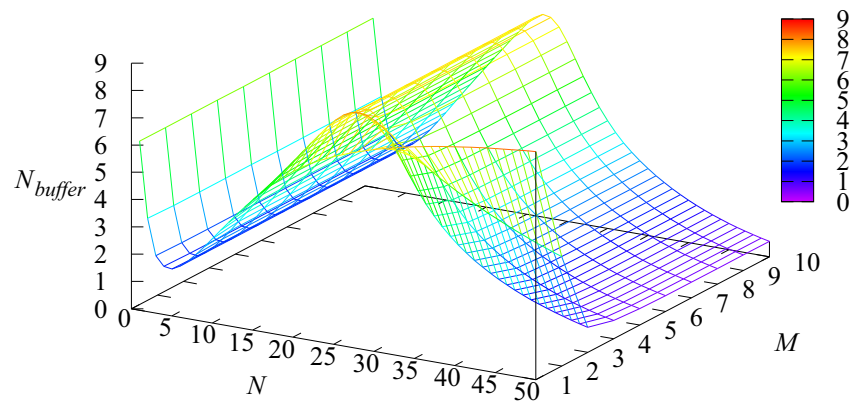


Figure 4. Dependence of the average number  $N_{buf}$  of customers in the buffer on  $N$  and  $M$ .

Figures 5–7 illustrate the dependencies of the average number  $N_{serv-1}$  of occupied servers, the average number  $N_{blocked-1}$  of blocked servers that are under unblocking by an assistant, and the average number  $N_{blocked-2}$  of blocked servers that are waiting until an assistant will become available on the parameters  $N$  and  $M$ .

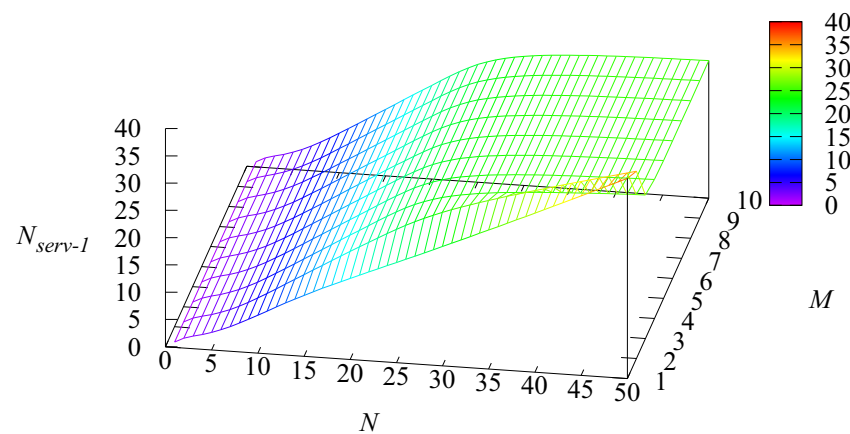


Figure 5. Dependence of the average number  $N_{serv-1}$  of occupied type-1 servers on  $N$  and  $M$ .

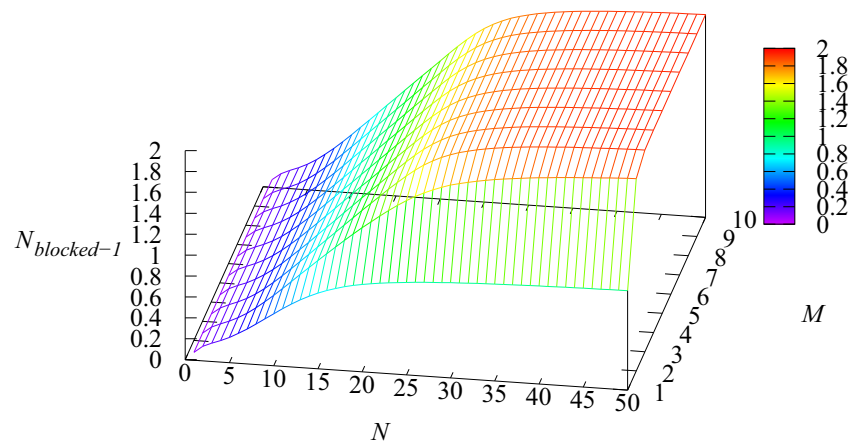


Figure 6. Dependence of the average number  $N_{blocked-1}$  on  $N$  and  $M$ .

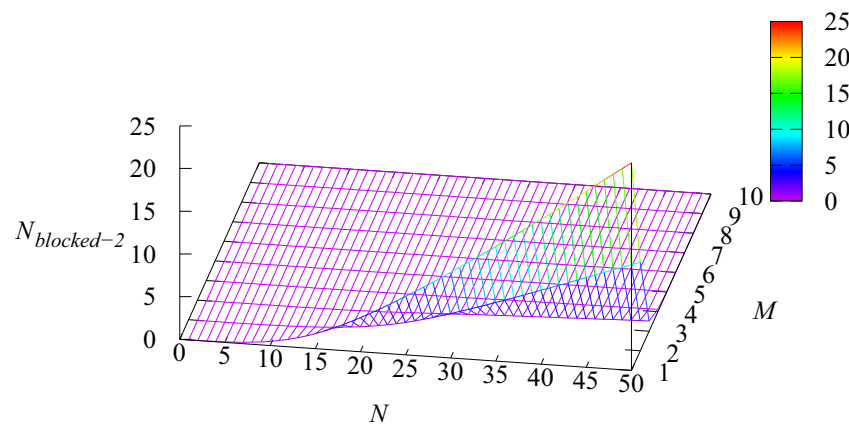


Figure 7. Dependence of the average number  $N_{blocked-2}$  on  $N$  and  $M$ .

The average number  $N_{serv-1}$  of occupied servers essentially increases with the increase in the total number of servers  $N$ . When the number of assistants  $M$  increases, the average number of blocked servers decreases if the average arrival rate is constant. Thus, the increase in the number of assistants  $M$  may lead to the decrease in the number of busy servers. However, with an increasing value of  $M$ , the rating and the intensity of the arrival flow can also increase, which may cause an increase in the number of busy servers. Therefore, the dependence of the number  $N_{serv-1}$  on the parameter  $M$  is hardly predictable.

The average number  $N_{blocked-1}$  of blocked servers currently receiving the help of assistants increases with the growth of  $N$  and  $M$ . This growth is mainly caused by the increase in the average arrival rate. In the considered example, the average number  $N_{blocked-2}$  of blocked servers that are waiting until an assistant will become available grows when the number of servers  $N$  grows and decreases with the increasing value of  $M$ .

The dependence of the loss probability  $P_{ent}$  of an arbitrary customer at the entrance to the system due to the unwillingness to wait in a long queue on the parameters  $N$  and  $M$  is presented in Figure 8.

As it is seen from Figure 8, the loss probability of an arbitrary customer at the entrance to the system decreases when the number  $N$  of servers grows. The dependence of  $P_{ent}$  on  $M$  is less essential, and  $P_{ent}$  may decrease or increase with the growth of  $M$ .

The dependence of the loss probability  $P_{imp}$  of an arbitrary customer due to impatience on the parameters  $N$  and  $M$  is presented in Figure 9.

The behavior of the dependence of the loss probability  $P_{imp}$  of an arbitrary customer due to impatience on the parameters  $N$  and  $M$  is also complicated. Note that the loss probability  $P_{imp}$  essentially depends on the average number of customers in the buffer (see Figure 4) and the average arrival rate  $\lambda$  (see Figure 3).

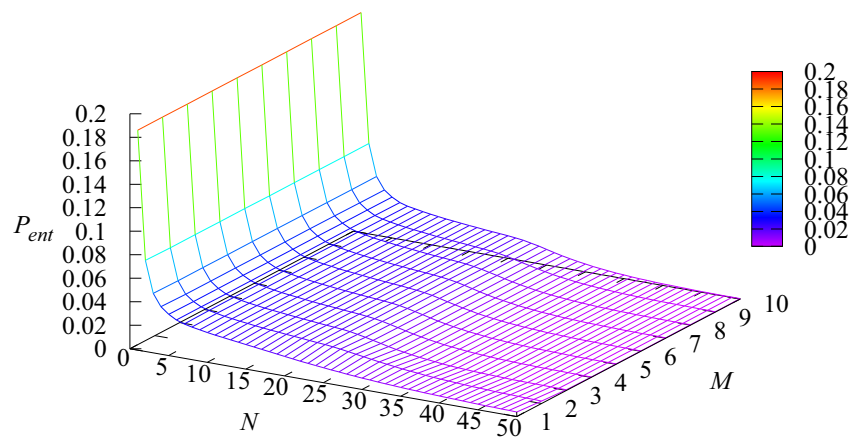


Figure 8. Dependence of the loss probability  $P_{ent}$  on  $N$  and  $M$ .

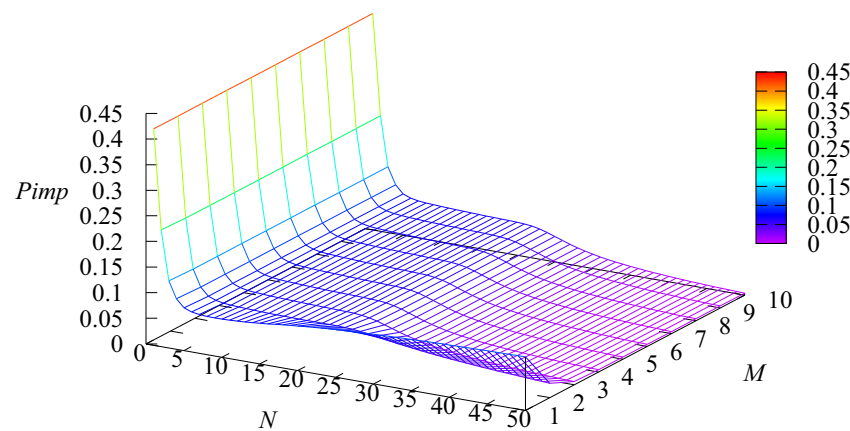


Figure 9. Dependence of the loss probability  $P_{imp}$  on  $N$  and  $M$ .

The dependence of the loss probability  $P_{loss}$  of an arbitrary customer on the parameters  $N$  and  $M$  is presented in Figure 10.

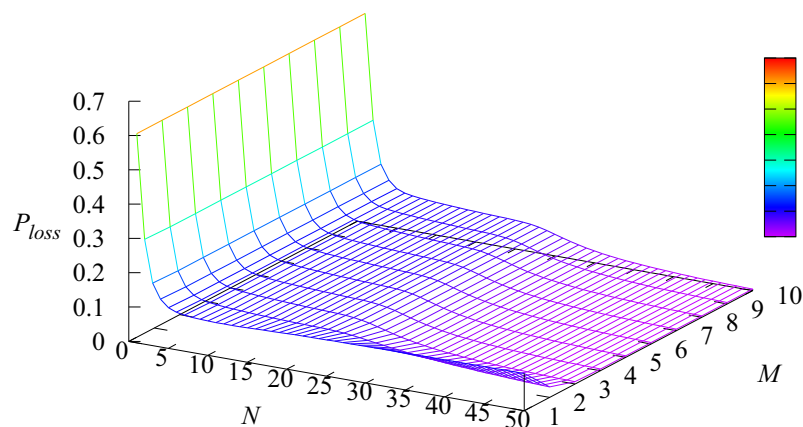


Figure 10. Dependence of the loss probability  $P_{loss}$  on  $N$  and  $M$ .

The loss probability  $P_{loss}$  is the sum of loss probabilities  $P_{ent}$  and  $P_{imp}$ . Because in the considered case, the main losses occur due to customers' impatience, the behavior of  $P_{loss}$  is similar to the behavior of the probability  $P_{imp}$ .

We considered the dependencies of the main performance measures on the number of servers  $N$  and the number of assistants  $M$  and can conclude that these dependencies can be quite hardly predictable. Therefore, mathematical modeling with the use of the results presented in this paper is required for the exact estimation of the system performance characteristics. It is worth noting that, from the point of view of potential practical

applications, the more important problem is to define the optimal number  $N$  of servers and the number  $M$  of assistants. To find these optimal values, first of all, it is necessary to choose an appropriate cost criterion. In this paper, we assume that the quality of the system operation is characterized by the following economic criterion:

$$E = E(M, N) = a_1\lambda_{out} - b_1\lambda P_{ent} - b_2\lambda P_{imp} - d_1N - d_2M.$$

Here, the cost coefficients  $a_1, b_1, b_2, d_1,$  and  $d_2$  have the following meaning:

$a_1$  is the profit obtained by the system for the successful service of one customer;

$b_1$  and  $b_2$  are the charges paid by the system for the loss of a customer at the entrance of the system and due to impatience, correspondingly;

$d_1$  and  $d_2$  are the charges paid by the system for maintaining one server and one assistant per unit time, correspondingly;

Thus, the economic criterion  $E$  has the meaning of the average profit obtained by the system per unit of time.

In this numerical example, we fix the following cost coefficients:

$$a_1 = 1, b_1 = 2, b_2 = 3, d_1 = 0.05, d_2 = 0.1.$$

We aim to maximize the average profit of the system.

Figure 11 illustrates the dependence of the economic criterion  $E$  on the parameters  $N$  and  $M$ .

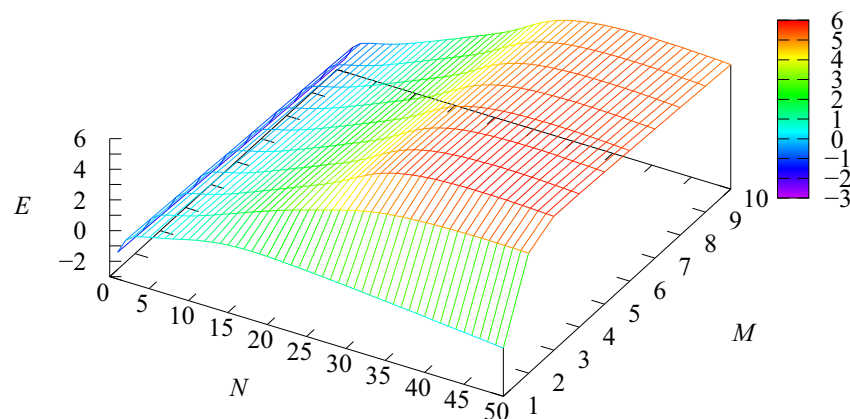


Figure 11. Dependence of the values of economic criterion  $E$  on  $N$  and  $M$ .

As is seen from Figure 11, the profit of the system essentially increases when the number of the exploited servers grows from 1 to about 30. When the number of servers grows from 30 to 40, the profit still increases, but not so essentially. A further increase in the number of servers  $N$  leads to a slight decrease in the system profit because the cost of adding a new server exceeds the obtained additional profit from the service of customers. The same situation occurs when the number  $M$  of assistants grows. Thus, it can be verified that in the considered example, the optimal value  $E^*$  of the economic criterion  $E$  is  $E^* = 5.87082$ . This optimal value is achieved when the number of servers is  $N = 40$ , and the number of assistants is  $M = 4$ .

It is worth noting that the maximal size of the blocks of the generator is defined by the number  $(N + 1)RW$ . One of the goals of this numerical experiment was to demonstrate the feasibility of the elaborated algorithms for more or less realistic values of the number  $N$  of SSDs and the number  $M$  of assistants. In this experiment, we have fixed  $N = 50$  and  $M = 10$  and the range of rating as the set  $\{1, \dots, 10\}$ , which is quite enough for the modeling of even a quite large real hypermarket. Therefore, the maximal size of the block in this example is 510  $W$ . Because the total number of points  $(N, M)$  for which computations were performed to show the shape of the considered dependencies and solve the optimization problem is  $NM = 500$ , we restricted ourselves by the base  $MAP$  of

order 2, just to avoid long computations. For  $W = 2$ , the maximal size of the block is more than 1000, and computation for 500 points takes several minutes. The increase in  $W$  does not create any essential problem in computations except the increase in computation time with the increase in  $W$ . Note that the use of the *MAP* of order 2 is often enough for good matching the main characteristics of the *MAP* flow to the corresponding characteristics of even quite bursty real flows.

## 8. Conclusions

In this paper, we have considered a queueing model having a finite number  $N$  of servers and  $M$  assistants that help the servers when certain service problems occur. This model fits a description of the operation of a huge variety of real-world systems with so-called self-service of customers. We assume the novel description of an arrival process as the generalization of the *MAP* to the case of rating-dependent arrival rates. Rating dependent arrivals are typical in many real systems with competing service providers. The effects of possible customers abandonment (balking) and impatience (reneging) are accepted for consideration. Algorithmic analysis of this queueing model based on the use of level-dependent multi-dimensional Markov chains is implemented. This analysis includes the derivation of conditions for stability and non-stability of the model, computation of the steady-state distribution of the chain, numerical illustration of the dependence of the main performance measures of the system on  $N$  and  $M$ , and the solution of the optimization problem.

The obtained results can be extended to the models with other possible mechanisms for calculating the value of the rating, more complicated distributions of service and help times, unreliable servers or (and) assistants, heterogeneous (experienced and non-experienced) customers, etc.

**Author Contributions:** Conceptualization, S.D., Y.G. and A.D.; methodology, S.D., O.D. and Y.G.; software, S.D. and O.D.; validation, S.D. and O.D.; formal analysis, S.D., Y.G. and A.D.; investigation, A.D.; writing, original draft preparation, Y.G. and A.D.; writing, review and editing A.D. and S.D.; supervision A.D. and Y.G.; project administration O.D. and A.D. All authors read and agreed to the published version of the manuscript.

**Funding:** This paper has been supported by the RUDN University Strategic Academic Leadership Program.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Huang, F.; Guo, P.; Wang, Y. Cyclic Pricing When Customers Queue with Rating Information. *Prod. Oper. Manag.* **2019**, *28*, 2471–2485. [[CrossRef](#)]
2. Dudin, S.; Dudin, A.; Dudina, O.; Samouylov, K. Competitive queueing systems with comparative rating dependent arrivals. *Oper. Res. Perspect.* **2020**, *7*, 100139. [[CrossRef](#)]
3. Meuter, M.L.; Ostrom, A.L.; Roundtree, R.I.; Bitner, M.J. SSTs: Understanding customer satisfaction with technology-based service encounters. *J. Mark.* **2000**, *54*, 50–64. [[CrossRef](#)]
4. Lyu, F.; Lim, H.A.; Choi, J. Customer Acceptance of Self-service Technologies in Retail: A Case of Convenience Stores in China. *Asia Pac. J. Inf. Syst.* **2019**, *29*, 428–447. [[CrossRef](#)]
5. Li, M.; Choi, T.Y.; Rabinovich, E.; Crawford, A. Self-service operations at retail stores: The role of inter-customer interactions. *Prod. Oper. Manag.* **2013**, *22*, 888–914. [[CrossRef](#)]
6. Kokkinou, A.; Cranage, D.A. Using self-service technology to reduce customer waiting times. *Int. J. Hosp. Manag.* **2013**, *33*, 435–445. [[CrossRef](#)]
7. Djelassi, S.; Diallo, M.F.; Zielke, S. How self-service technology experience evaluation affects waiting time and customer satisfaction? A moderated mediation model. *Decis. Support Syst.* **2018**, *111*, 38–47. [[CrossRef](#)]



8. Hwang, Y.; Kim, D.J. Customer self-service systems: The effects of perceived Web quality with service contents on enjoyment, anxiety, and e-trust. *Decis. Support Syst.* **2007**, *43*, 746–760. [[CrossRef](#)]
9. Morimura, F.; Nishioka, K. Waiting in exit-stage operations: expectation for self-checkout systems and overall satisfaction. *J. Mark. Channels* **2016**, *23*, 241–254. [[CrossRef](#)]
10. Wu, C.H.; Yang, D.Y. Bi-objective optimization of a queueing model with two-phase heterogeneous service. *Comput. Oper. Res.* **2021**, *130*, 105230. [[CrossRef](#)]
11. Neuts, M. *Matrix-Geometric Solutions in Stochastic Models*; John Hopkins University Press: Baltimore, MD, USA, 1981.
12. Haque, L.; Armstrong, M.J. A survey of the machine interference problem. *Eur. J. Oper. Res.* **2007**, *179*, 469–482. [[CrossRef](#)]
13. Singh, P.R.; Sharma, G.C.; Jain, M. Some perspectives of machine repair problems. *Int. J. Eng.* **2010**, *23*, 253–268.
14. Chakravarthy, S.R. A multi-server queueing model with server consultations. *Eur. J. Oper. Res.* **2014**, *233*, 625–639. [[CrossRef](#)]
15. Lucantoni, D.M. New Results on the Single Server Queue with a Batch Markovian Arrival Process. *Commun. Stat. Stoch. Model.* **1991**, *7*, 1–46. [[CrossRef](#)]
16. Chakravarthy, S.R. The batch Markovian arrival process: A review and future work. *Adv. Probab. Theory Stoch. Process.* **2001**, *1*, 21–49.
17. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer Nature: Cham, Switzerland, 2019; ISBN 978-3-030-32072-0.
18. Vishnevskii, V.M.; Dudin, A.N. Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. *Autom. Remote Control* **2017**, *78*, 1361–1403. [[CrossRef](#)]
19. Krishnamoorthy, A.; Thekkiniyedath, R.; Lakshmy, B. On a Two-Server Queue with Consultation in Random Environment. In *International Conference on Information Technologies and Mathematical Modelling*; Springer: Cham, Switzerland, 2019; pp. 217–229.
20. Resmi, T.; Ravikumar, K. Three-Server Queue with Consultations by Main Server with a Buffer at the Main Server. In *International Conference on Information Technologies and Mathematical Modelling*; Springer: Cham, Switzerland, 2020; pp. 131–142.
21. Lal, T.S.; Krishnamoorthy, A.; Joshua, V.C.; Vishnevsky, V. A Two-Stage Tandem Queue with Specialist Servers. In *Applied Probability and Stochastic Processes*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 335–353.
22. Klimenok, V.; Dudin, A.; Samouylov, K. Analysis of the  $BMAP/PH/N$  queueing system with backup servers. *Appl. Math. Model.* **2018**, *57*, 64–84. [[CrossRef](#)]
23. Xu, S.H.; Gao, L.; Ou, J. Service performance analysis and improvement for a ticket queue with balking customers. *Manag. Sci.* **2007**, *53*, 971–990. [[CrossRef](#)]
24. Batt, R.J.; Terwiesch, C. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Manag. Sci.* **2015**, *61*, 39–59. [[CrossRef](#)]
25. Sun, B.; Dudin, A.; Dudin, S. Queueing system with impatient customers, visible queue and replenishable inventory. *Appl. Comput. Math.* **2018**, *17*, 161–174.
26. Wang, K.; Li, N.; Jiang, Z. Queueing system with impatient customers: A review. In Proceedings of the 2010 IEEE International Conference on Service Operations and Logistics, and Informatics, Qingdao, China, 15–17 July 2010; pp. 82–87.
27. Hanukov, G.; Hassoun, M.; Musicant, O. On the Benefits of Providing Timely Information in Ticket Queues with Balking and Calling Times. *Mathematics* **2021**, *9*, 2753. [[CrossRef](#)]
28. Jouini, O.; Koole, G.; Roubos, A. Performance indicators for call centers with impatient customers. *IIE Trans.* **2013**, *45*, 341–354. [[CrossRef](#)]
29. Garnett, O.; Mandelbaum, A.; Reiman, M. Designing a call center with impatient customers. *Manuf. Serv. Oper. Manag.* **2002**, *4*, 208–227. [[CrossRef](#)]
30. Boots, N.K.; Tijms, H. A multiserver queueing system with impatient customers. *Manag. Sci.* **1999**, *45*, 444–448. [[CrossRef](#)]
31. De Kok, A.G.; Tijms, H.C. A queueing system with impatient customers. *J. Appl. Probab.* **1985**, *22*, 688–696. [[CrossRef](#)]
32. Boots, N.K.; Tijms, H. An  $M/M/c$  queue with impatient customers. *Top* **1999**, *7*, 213–220. [[CrossRef](#)]
33. Shin, Y.W.; Choo, T.S.  $M/M/s$  queue with impatient customers and retrials. *Appl. Math. Model.* **2009**, *33*, 2596–2606. [[CrossRef](#)]
34. Kuki, A.; Berczes, T.; Sztrik, J.; Toth, A. Reliability analysis of a retrial queueing systems with collisions, impatient customers, and catastrophic breakdowns. In Proceedings of the 2021 International Conference on Information and Digital Technologies (IDT), Zilina, Slovakia, 22–24 June 2021; pp. 254–259.
35. Wüchner, P.; Sztrik, J.; De Meer, H. Finite-source  $M/M/S$  retrial queue with search for balking and impatient customers from the orbit. *Comput. Netw.* **2009**, *53*, 1264–1273. [[CrossRef](#)]
36. Kuki, A.; Berczes, T.; Toth, A.; Sztrik, J. Numerical analysis of finite source Markov retrial system with non-reliable server, collision, and impatient customers. *Ann. Math. Inform.* **2020**, *51*, 53–63. [[CrossRef](#)]
37. Klimenok, V.I.; Orlovsky, D.S.; Dudin, A.N. A  $MAP/PH/N$  system with impatient repeated calls. *Asia-Pac. J. Oper. Res.* **2007**, *24*, 293–312. [[CrossRef](#)]
38. Van Do, T.; Do, N.H.; Zhang, J. An enhanced algorithm to solve multiserver retrial queueing systems with impatient customers. *Comput. Ind. Eng.* **2013**, *65*, 719–728.
39. Danilyuk, E.; Vygoskaya, O.; Moiseeva, S. Retrial queue  $M/M/N$  with impatient customer in the orbit. In Proceedings of the International Conference on Distributed Computer and Communication Networks, Moscow, Russia, 17–21 September 2018; pp. 493–504.
40. Perel, N.; Yechiali, U. Queues with slow servers and impatient customers. *Eur. J. Oper. Res.* **2010**, *201*, 247–258. [[CrossRef](#)]



41. Dudin, S.A.; Dudina, O.S. Call center operation model as a  $MAP/PH/N/R - N$  system with impatient customers. *Probl. Inf. Transm.* **2011**, *47*, 364–377. [[CrossRef](#)]
42. Kim, C.; Dudin, A.; Dudin, S.; Dudina, O. Tandem queueing system with impatient customers as a model of call center with Interactive Voice Response. *Perform. Eval.* **2013**, *70*, 440–453. [[CrossRef](#)]
43. Dudin, S.; Kim, C.; Dudina, O.  $MMAP/M/N$  queueing system with impatient heterogeneous customers as a model of a contact center. *Comput. Oper. Res.* **2013**, *40*, 1790–1803. [[CrossRef](#)]
44. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Courier Dover Publications: New York, NY, USA, 2018.
45. Klimenok, V.I.; Dudin, A.N. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Syst.* **2006**, *54*, 245–259. [[CrossRef](#)]
46. Dudin, S.; Dudina, O. Retrial multi-server queueing system with PHF service time distribution as a model of a channel with unreliable transmission of information. *Appl. Math. Model.* **2019**, *65*, 676–695. [[CrossRef](#)]
47. Dudin, S.; Dudin, A.; Kostyukova, O.; Dudina, O. Effective algorithm for computation of the stationary distribution of multi-dimensional level-dependent Markov chains with upper block-Hessenberg structure of the generator. *J. Comput. Appl. Math.* **2020**, *366*, 112425. [[CrossRef](#)]