mathematics

MDPI

Article

Analysis of Multi-Server Queue with Self-Sustained Servers

Alexander Dudin [1,2,*], Olga Dudina [1], Sergei Dudin [1] and Konstantin Samouylov [2]

[1] 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.)
[2] Applied Mathematics and Communications Technology Institute, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St., 117198 Moscow, Russia; ksam@sci.pfu.edu.ru
* Correspondence: dudin@bsu.by or dudin_alexander@mail.ru

Abstract: A novel multi-server vacation queuing model is considered. The distinguishing feature of the model, compared to the standard queues, is the self-sufficiency of servers. A server can terminate service and go on vacation independently of the system manager and the overall situation in the system. The system manager can make decisions whether to allow the server to start work after vacation completion and when to try returning some server from a vacation to process customers. The arrival flow is defined by a general batch Markov arrival process. The problem of optimal choice of the total number of servers and the thresholds defining decisions of the manager arises. To solve this problem, the behavior of the system is described by the three-dimensional Markov chain with the special block structure of the generator. Conditions for the ergodicity of this chain are derived, the problem of computation of the steady-state distribution of the chain is discussed. Expressions for the key performance indicators of the system in terms of the distribution of the chain states are derived. An illustrative numerical result is presented.

Keywords: multi-server vacation queuing model; self-sufficient servers; optimization; multi-dimensional Markov chains

Citation: Dudin, A.; Dudina, O.; Dudin, S.; Samouylov, K. Analysis of Multi-Server Queue with Self-Sustained Servers. *Mathematics* **2021**, *9*, 2134. https://doi.org/10.3390/math9172134

Academic Editors: Ivan Atencia and José Luis Galán-García

Received: 12 August 2021
Accepted: 31 August 2021
Published: 2 September 2021

1. Introduction

Queuing theory is one of the most quickly developing branches of applied probability due to the intensive appearance of new queuing models of real-world systems, telecommunication networks, in particular. As a consequence of the creation of various new schemes of resource sharing and multiplexing, the role of multi-server systems grows. The most common assumption made in the literature about multi-server queues is that there exists a finite or infinite pool of servers that provide service to arriving customers under the control of the system manager. The simplest, traditional, case assumes the existence of the fixed finite number of servers. The arriving customer occupies one of the idle servers to receive service. If all servers are occupied, the arrived customer is lost or stored into the buffer of a finite or infinite capacity or temporarily leaves the system and makes retrials to find a free server.

The process of servers' occupation is usually completely defined by the arrival of customers. The arrival of a new customer implies the occupation of one more server if some server is free. The traditional service disciplines for the systems with a buffer are so-called conservative. This means that the situation when the input buffer is not empty while some servers are idle is not possible. The extreme diversity of real-world systems and processes that can be modeled by multi-server queuing systems gave raise to consideration of other scenarios of involvement of the servers into the processing of customers. Some of them are briefly listed below.

- Systems with unreliable servers in which any server (or all servers together) can be broken after a random period of time and require recovering. During the server recovering, it is not able to provide service. After the end of the recovering, the server

resumes service if the input buffer is not empty. The literature about unreliable queuing systems (or queues with service interruption) is huge. We can mention, e.g., [1–10].

- Systems with server vacation in which in some situations the server can temporally stop operation and go to so-called vacation. Usually, the decision of whether or not to go on vacation does not depend on the wish of the server. Vacation is automatically taken if the buffer is empty or restriction on the continuous time of operation of the server or the number of sequentially provided services is violated. The existing literature about queuing systems with server vacation is also huge. We can mention, e.g., [11–16]; however, it is worth noting that the research of queuing systems with vacations mainly focuses on the analysis of single-server queues, e.g., [17].
- Systems with randomly varying number of available servers as the partial case of queues operating in the random environment, see, e.g., [18,19] and references therein. Note that in the majority of the papers devoted to the queues operating in the random environment the number of the servers does not change. Only the arrival, service and other rates are changed at the epochs of the change of the state of the random environment. These systems can be regarded as the generalization of unreliable queues to the cases of more flexible mechanisms of server activation and deactivation as well as partial reliability of the servers. The study of systems operating in a random environment was apparently initiated in [20] under the name of systems with partial failure of the device. Links to numerous more recent publications can be found, e.g., in [10,21–23].
- Systems with a controllable number of active servers in which a certain part of existing servers is switched-off and is activated (one-by-one or in blocks) only when the number of customers in the system grows above some fixed threshold values, see, e.g., [24–27].
- Systems in which arriving customers require occupation of a random number of servers, see, e.g., [28], or when all servers not engaged currently into service have to start service of an arriving customer, see, e.g., [29,30].
- Systems with additional resource required for service, see, e.g., [31–36]. Such systems are called in the literature as queuing-inventory systems, assembly-like systems, double-sided or queues with paired customers. These systems are non-conservative because the servers can stay idle in the presence of a non-empty buffer if some additional resource required for service is currently unavailable.

The model considered in this paper combines certain features of systems with vacations and systems with a controllable number of active servers. The common features are an opportunity of taking a vacation by a server and the possibility of influence, to some extent, on these vacations. The distinguishing feature of the considered model is the self-sufficiency of the servers. This self-sufficiency manifests itself as follows.

(i) Any server that just finished service of a customer may decide to take a vacation or start a new service, independently of the situation in the system and desire of the system manager;

(ii) The server being on vacation (we refer to such a server as the vacated server) that obtains an invitation (offer) of the system manager to terminate a vacation, due to accumulation of a long queue, can accept this offer and start service or decline this offer and continue its vacation.

Note that the leverage of influence of the system manager on the servers is the option not to allow to finish a vacation and start service if the queue length in the system is small.

The goal of this paper is to apply the matrix-analytical methods of queuing theory for analysis of the queuing system with self-sustained servers, which have many important applications. Analysis is far from being easy due to the space-inhomogeneous behavior of the multi-dimensional Markov chain describing the dynamics of the system. Nevertheless, the stability condition of the system is obtained in a nice analytically tractable form,

algorithms for computation of the main performance measures of the system are elaborated and numerically verified.

The novelty of this paper is defined by the fact that, to the best of our knowledge, the systems with the self-sufficiency of servers defined in this paper are not considered in the existing literature.

One of the numerous possible practical applications of the considered queuing model is as follows. Let us consider a manufacturing company that produces some product. Products arrive at the warehouse from where they need to be delivered to consumers by trucks. The business model of the company does not assume the purchase, storage and maintenance of trucks. Instead, the company hires some pool of independent individual cargo carriers (freelancers, individual entrepreneurs) having their own truck for delivering the products. The contract between the company and the truck driver (server) assumes the obligation of the driver to finish any started service and its right not to provide further services for a while. The obligation of the company may be providing an opportunity to each server to implement a certain minimum average number of services during a fixed period of time. The company has the right to ignore the wish of the server to provide a service (if the queue is empty or short) and to invite the vacated server for resuming the service (if the queue becomes long).

A similar system arises in modeling Internet-based taxi companies in which the majority of engaged taxi drivers work only part-time in comfortable periods of time for them.

The proposed model can be used by the company owner for managerial goals. Namely, he/she can decide the optimal number of contracts with servers to be signed, conditions of proposing to the server to interrupt vacation and declining the attempt of the server to start working after vacation completion. Various costs can be taken into account, e.g., average waiting time in the queue, queue length, probability of a customer loss due to impatience, and charge for not providing the chances to implement the negotiated number of services. The problem of optimization of the number of required servers is far from trivial. If this number is too small, a queue may be pretty high. Customers may leave the queue without service (if the products are perishable or must be delivered to the consumer by a certain time), which may provoke substantial losses for the company. If this number is too large, problems with providing enough work to the servers can arise. The solution of the optimization problem requires the computation of the values of the various performance indicators of the system under the fixed values of the system parameters and the thresholds of defining decisions about sending invitations to the vacated servers and declining their offers to start working. As the mandatory step to make this computation, it is necessary to describe the behavior of the system by the multi-dimensional Markov chain and analyze its stationary distribution. Taking into account the nice properties of the Kronecker product of matrices, it is possible to write down the generator of the Markov chain describing the operation of the transportation company in a transparent form. Then, it is necessary to analyze the stationary behavior of the chain, compute performance measures of the system and implement numerical analysis of these indicators.

Correspondingly, the outline of the content of the sections of the paper is as follows. Section 2 contains a precise description of the queuing system under study. The process describing the behavior of the system is defined in Section 3 as the three-dimensional Markov chain. The infinitesimal generator of this Markov chain is written down. Section 4 is devoted to the derivation of the ergodicity condition of this Markov chain. Section 5 concerns the problem of computation of the stationary distribution of this Markov chain. Formulas for computing the values of several performance indicators of the system based on the known stationary distribution of the Markov chain are presented in Section 6. Section 7 contains numerical illustrations and Section 8 concludes the paper.

2. Mathematical Model

We consider a queuing system having the structure presented in Figure 1.

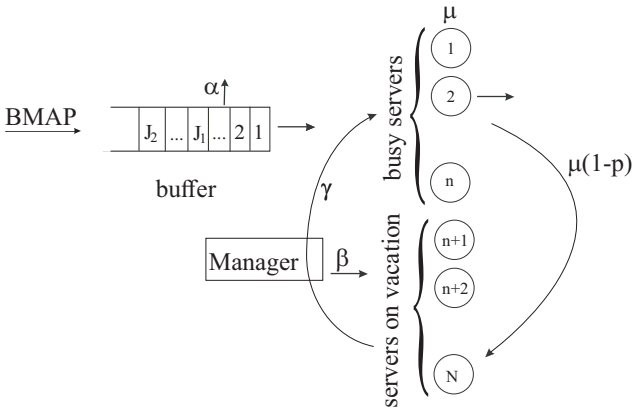

Figure 1. Structure of queuing system under study.

The system has N identical and independent of each other servers. All arriving customers are temporarily placed into the buffer of an infinite capacity. The service discipline is First-In-First-Out. The arrival of customers is described by the batch Markov arrival process ($BMAP$). The $BMAP$ is a rich class of point processes that allows to adequately model the majority of real-world processes, including those having correlated inter-arrival times and (or) large variance of these times. This makes them valuable for modeling real-world arrival flows. A detailed description of the $BMAP$ can be found, e.g., in [37–39]; therefore, we present only very dense information about the $BMAP$ that is necessary for the goals of this paper.

Arrivals of customers in the $BMAP$ occur under control of an irreducible Markov chain ν_t, $t \geq 0$, with a finite state space $\{1, \ldots, W\}$ were W is the fixed integer number. Intensities of transitions of the process ν_t, which are accompanied by the arrival of a batch of k customers, are given by the entries of the matrix D_k, $k \geq 0$. Here we assume that $D_k = O$ for $k > K$ where K is the maximum batch size. This assumption is just a technical one. It is non-necessary and made just for receiving a possibility to use for further analysis the numerically more efficient algorithms from [40] along with the algorithms from [41].

The diagonal entries of the matrix D_0 are negative and define, up to the sign, the intensity of departing of process ν_t from the corresponding states. The matrix $D = \sum_{k=0}^{K} D_k$ is the generator of Markov chain ν_t. The stationary distribution of Markov chain ν_t is defined by the row vector $\boldsymbol{\theta}$ that is the unique solution to the system

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

Here and in the sequel, $\mathbf{e}$ is a column vector of appropriate size consisting of 1 s, $\mathbf{0}$ is a row vector of appropriate size consisting of 0 s.

The average rate λ of customers arrival is defined by

$$\lambda = \boldsymbol{\theta} \sum_{k=0}^{K} k D_k \mathbf{e}.$$

The average rate λ_b of batches arrival is defined by

$$\lambda_b = \boldsymbol{\theta} \sum_{k=0}^{K} D_k \mathbf{e}.$$

We assume that the service time of a customer is exponentially distributed with the parameter μ, $\mu < \infty$. If during the service completion epoch at some server the buffer is idle, the server takes a vacation. If the buffer is not idle, the server starts a new service with probability p, $0 \leq p \leq 1$, and with the complementary probability takes a vacation independently on the number of customers in the buffer. We assume that the vacation

times of different servers are independent and have an exponential distribution with the parameter γ, $0 < \gamma < \infty$. When the vacation time of a server expires, the server tries to start work. If during the vacation completion moment the number of customers in the buffer is greater than or equal to the predefined threshold $J_1, 0 < J_1 < \infty$, the server picks up the customer from the head of the queue and starts its service. In the opposite case, i.e., the queue length is less than J_1, the server has to repeat vacations until it will obtain permission to resume service.

We assume that when the number of customers in the buffer becomes greater or equal to another predefined threshold J_2 such that $J_1 < J_2 < \infty$, the system starts searching a server being on vacation and offers to the server to terminate a vacation ahead of the schedule and start service. The search time has the exponential distribution with the parameter β, $0 < \beta < \infty$. During each search, the system offers to start working to vacated servers sequentially. If any server agrees to start work, the search is stopped. If a server declines the offer, the system makes an offer to another vacated server. We assume that each vacated server declines the offer with probability q, $0 \leq q \leq 1$, and accepts the offer with the complementary probability. Thus, with the probability q^n, where n is the number of servers on the vacation, no one vacated server wants to interrupt the vacation and start working. With the complementary probability, the search results function by finding the server who accepts the offer. This server starts to work immediately. After the end of any search, the system starts the new search if the number of customers in the buffer is still greater or equal to the threshold J_2. Otherwise, new searches are not implemented until the queue length increases again to the level J_2.

Customers can be impatient and leave the buffer without service, independently of other customers. A customer leaves the system without service after an exponentially distributed patience time with the parameter α, $\alpha \geq 0$.

Our goal is to analyze the described queuing 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

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

where, during the epoch t,

- i_t is the number of customers in the buffer, $i_t \geq 0$;
- n_t is the number of busy servers, $n_t = \overline{0, N}$;
- ν_t is the state of the underlying process of the $BMAP$, $\nu_t = \overline{1, W}$.

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

To formally define the continuous-time Markov ξ_t, it is necessary to write down, for any pair of the states (i, n, ν) and (i', n', ν'), the intensities of transition 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 $\xi_t = \{i_t, n_t, \nu_t\}$ in the direct lexicographic order of the components $\{n_t, \nu_t\}$ and combine the set of states $(i, 0, 1), \ldots, (i, 0, W), (i, 1, 1), \ldots, (i, 1, W), \ldots, (i, N, 1), \ldots, (i, N, W)$ into the *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 ξ_t, $t \geq 0$.

Theorem 1. *The generator Q of the Markov chain $\xi_t,\ t \geq 0$, has the following block upper-Hessenbergian structure*

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

The non-zero blocks are defined as follows:

$$Q_{0,0} = I_{N+1} \otimes D_0 - \mu C \otimes I_W + \mu C E^- \otimes I_W,$$

$$Q_{i,i} = I_{N+1} \otimes D_0 - i\alpha I_{(N+1)W} - \mu C \otimes I_W + \mu(1-p) C E^- \otimes I_W,\ 0 < i < J_1,$$

$$Q_{i,i} = I_{N+1} \otimes D_0 - i\alpha I_{(N+1)W} - \mu C \otimes I_W + \mu(1-p) C E^- \otimes I_W -$$
$$-\gamma(N I_{N+1} - C) \otimes I_W,\ J_1 \leq i < J_2,$$

$$Q_{i,i} = I_{N+1} \otimes D_0 - i\alpha I_{(N+1)W} - \mu C \otimes I_W + \mu(1-p) C E^- \otimes I_W -$$
$$-\gamma(N I_{N+1} - C) \otimes I_W - \beta(I_{N+1} - \tilde{Q}) \otimes I_W,\ i \geq J_2,$$

$$Q_{i,i+k} = I_{N+1} \otimes D_k,\ k = \overline{1,K},$$

$$Q_{i,i-1} = i\alpha I_{(N+1)W} + \mu p C \otimes I_W,\ 0 < i \leq J_1,$$

$$Q_{i,i-1} = i\alpha I_{(N+1)W} + \mu p C \otimes I_W + \gamma(N I_{N+1} - C) E^+ \otimes I_W,\ J_1 \leq i < J_2,$$

$$Q_{i,i-1} = i\alpha I_{(N+1)W} + \mu p C \otimes I_W + \gamma(N I_{N+1} - C) E^+ \otimes I_W + \beta(I_{N+1} - \tilde{Q}) E^+ \otimes I_W,\ i \geq J_2,$$

where

$\otimes$ *indicates the symbol of the Kronecker product matrices, see [42];*

$C = \mathrm{diag}\{0, 1, \dots, N\}$;

$\mathrm{diag}\{\dots\}$ *denotes the diagonal matrix with the diagonal entries listed in the brackets;*

$I_{\cdot}$ *is the identity matrix having size indicated in the suffix (if the size of the matrix is clear from the context, it can be omitted);*

$$E^- = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix},\ E^+ = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix};$$

$\tilde{Q} = \mathrm{diag}\{q^N, q^{N-1}, q^{N-2}, \dots, , q^1, 0\}$.

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

The diagonal entries of the matrix Q are negative. The modulus of each diagonal entry defines the intensity of leaving the corresponding state of the Markov chain $\xi_t,\ t \geq 0$. If the number of customers in the buffer is equal to zero, the Markov chain $\xi_t,\ t \geq 0$, can leave its state only if the underlying process of the $BMAP$ makes a transition without generation of a customer (the intensities of such transitions are given by the diagonal entries of the matrix D_0) or the service of a customer is finished (the intensities of such events are given as the diagonal entries of the matrix $\mu C \otimes I_W$). When the number of customers in the buffer is more than zero, additionally the Markov chain $\xi_t,\ t \geq 0$, can leave its state due to the departure of a customer from the buffer due to impatience (the intensities of such events are given as the diagonal entries of the matrix $i\alpha I_{(N+1)W}$). If the number of customers in the buffer is greater than or equal to J_1, the Markov chain ξ_t can also change its state

when a vacant server decides to start work (the intensities of such events are given as the diagonal entries of the matrix $\gamma(NI_{N+1} - C) \otimes I_W$). If the number of customers in the buffer is greater than or equal to J_2, the system can find a vacant server who agrees to start working, which also leads to the change of the state of the chain ξ_t (the intensities of such events are given as the diagonal entries of the matrix $\beta(I_{N+1} - \tilde{Q}) \otimes I_W$).

The non-diagonal entries of the matrix $Q_{i,i}$ define the intensities of the transition of the chain ξ_t that do not lead to the change of the number of customers in the buffer. In the case $i = 0$, such intensities are the intensity of service completion in one busy server (such intensities are the entries of the matrix $\mu CE^- \otimes I_W$). In the case $i > 0$, such intensities are the intensity of service completion and unwillingness to continue the work in one busy server (such intensities are the entries of the matrix $\mu(1-p)CE^- \otimes I_W$). Note, that in these cases the number of busy servers decreases by one. This explains the appearance of the matrix E^- in the corresponding formulas. Further, the transitions of the $BMAP$ underlying process without generation of batches of customers (the intensities of such transitions are given as the non-diagonal entries of the matrix D_0) does not imply the change of the number of customers in the buffer.

Taking into account all the above reasonings, we obtain the form of the non-diagonal blocks $Q_{i,i}$, $i \geq 0$.

The entries of the matrix $Q_{i,i-1}$ define the intensities of the Markov chain ξ_t transitions that lead to the decrease in the number of customers in the buffer by one. In the case $i \geq J_2$, the events that lead to the decreasing the number of customers in the buffer by one are the following:

1. The server finishes service and wants to continue the work (the corresponding intensities are given as the entries of the matrix $\mu pC \otimes I_W$);
2. One of the customers staying in the buffer abandons it due to impatience (the corresponding intensities are given as the entries of the matrix $i\alpha I_{(N+1)W}$);
3. One of the vacant servers decides to start working (the corresponding intensities are given as the entries of the matrix $\gamma(NI_{N+1} - C)E^+ \otimes I_W$);
4. The system finds a vacant server who agrees to work (the corresponding intensities are given as the entries of the matrix $\beta(I_{N+1} - \tilde{Q})E^+ \otimes I_W$);

Note that the occurrence of events 3 and 4 leads to an increase in the number of working servers by one, which is indicated by the matrix E^+ in the corresponding formulas. In the case $0 < i \leq J_1$, only events 1 and 2 are possible. In the case $J_1 \leq i < J_2$, possible events are 1, 2 and 3.

Taking these reasons into account we obtain the presented above form of the blocks $Q_{i,i-1}$, $i \geq 1$.

The entries of the matrix $Q_{i,i+k}$ define the intensities of the Markov chain ξ_t transitions that lead to the increase in the number of customers in the buffer by k, $k = \overline{1,K}$. This can happen only in the case of the arrival of a batch of customers of size k. The corresponding intensities are given as the entries of the matrices $I_{N+1} \otimes D_k$, which explains the form of the blocks $Q_{i,i+k}$, $i \geq 0$, $k = \overline{1,K}$. □

4. Ergodicity Condition

Before computing the stationary distribution of the Markov chain ξ_t, it is necessary to find the conditions that have to be imposed on the parameters of the system to guarantee the existence of this distribution (ergodicity conditions).

Theorem 2. *If the customers in the buffer are impatient, i.e., $\alpha > 0$, then the stationary distribution of the considered Markov chain ξ_t exists for any set of the system parameters.*

If the customers in the buffer are patient, i.e., $\alpha = 0$, then the necessary and sufficient condition of the existence of the stationary distribution of the considered Markov chain is defined by the inequality

$$\lambda < \sum_{n=0}^{N} x_n[np\mu + (N-n)\gamma + \beta(1-q^{N-n})] \quad (1)$$

where x_n is the probability that at an arbitrary moment when the system is overloaded, i.e., the number of customers in the buffer is huge, the number of busy servers is equal to n, $n = \overline{0, N}$.

The probabilities x_n are defined by formulas:

$$x_n = x_0 \prod_{l=1}^{n} \frac{\delta_{l-1}}{\epsilon_l},\ n = \overline{1, N}, \tag{2}$$

$$x_0 = \left(1 + \sum_{n=1}^{N} \prod_{l=1}^{n} \frac{\delta_{l-1}}{\epsilon_l}\right)^{-1} \tag{3}$$

where

$$\delta_n = \gamma(N - n) + \beta(1 - q^{N-n}),\ n = \overline{0, N-1}, \tag{4}$$

$$\epsilon_n = n(1 - p)\mu,\ n = \overline{1, N}. \tag{5}$$

Proof. Let us firstly consider the case $\alpha > 0$. In this case, the Markov chain ξ_t belongs to the class of asymptotically quasi-Toeplitz Markov chains, see [41]. This follows from the fact that the following limits exist:

$$Y^{(k)} = \lim_{i \to \infty} R_i^{-1} Q_{i,i+k-1} + \delta_{k,1} I,\ k = \overline{0, K+1}, \tag{6}$$

where R_i is the diagonal matrix with the diagonal entries equal to the corresponding diagonal entries of the matrix $Q_{i,i}$, $i \geq 0$, taken with the opposite sign, and $\delta_{k,1}$ is the Kronecker delta.

Let the matrix Y be defined as $Y = \sum_{k=0}^{K} Y^{(k)}$. The matrix Y can be reducible. We denote as Y_l the irreducible blocks of the canonical normal form of the matrix Y, and let $Y_l^{(k)}$ be the corresponding blocks of the matrices $Y^{(k)}$, $k = \overline{0, K+1}$.

For the ergodicity of an asymptotically quasi-Toeplitz Markov chain, it is required that for each irreducible block Y_l the following inequality

$$\mathbf{y}_l Y_l^{(0)} \mathbf{e} > \mathbf{y}_l \sum_{k=1}^{K} k Y_l^{(k+1)} \mathbf{e} \tag{7}$$

holds true.

Here, the vectors $\mathbf{y}_l$ are the unique solutions to the equations

$$\mathbf{y}_l Y_l = \mathbf{y}_l,\ \mathbf{y}_l \mathbf{e} = 1. \tag{8}$$

One can verify that in the considered case the matrix $Y^{(0)} = I_{(N+1)W}$ and the matrices $Y^{(k)} = O$, $k = \overline{1, K+1}$. Thus, ergodicity conditions (7) is transformed to the inequalities $1 > 0$ which are true for all possible system parameters.

Now, let us consider the case $\alpha = 0$. In this case, the matrices $Q_{i,i-1}$, $Q_{i,i}$ and $Q_{i,i+k}$, $i \geq J_2$, $k = \overline{1, K}$, do not depend on the value i and are defined as

$$Q_{i,i-1} = Q_0 = \mu p C \otimes I_W + \gamma(N I_{N+1} - C) E^+ \otimes I_W + \beta(I_{N+1} - \tilde{Q}) E^+ \otimes I_W,\ i \geq J_2,$$

$$Q_{i,i} = Q_1 = I_{N+1} \otimes D_0 - \mu C \otimes I_W + \mu(1 - p) C E^- \otimes I_W -$$
$$-\gamma(N I_{N+1} - C) \otimes I_W - \beta(I_{N+1} - \tilde{Q}) \otimes I_W,\ i \geq J_2,$$

$$Q_{i,i+k} = Q_{k+1} = I_{N+1} \otimes D_k,\ k = \overline{1, K}.$$

Thus, the Markov chain ξ_t belongs to the class of quasi-Toeplitz (or $M/G/1$ type) Markov chains. The sufficient and necessary condition for the ergodicity condition of a quasi-Toeplitz Markov chain, see, e.g., [39], is the fulfillment of the following inequality:

$$\mathbf{y}\sum_{k=1}^{K+1} kQ_k\mathbf{e} < 0$$

or

$$\mathbf{y}Q_0\mathbf{e} > \mathbf{y}\sum_{k=1}^{K} kQ_{k+1}\mathbf{e}. \tag{9}$$

Here, the vector $\mathbf{y}$ is the unique solution of the system

$$\mathbf{y}\hat{Q} = \mathbf{0},\ \mathbf{y}\mathbf{e} = 1, \tag{10}$$

where $\hat{Q} = \sum_{k=0}^{K+1} Q_k$.

It is easy to verify that the matrix $\hat{Q}$ has the form

$$\hat{Q} = I_{N+1} \otimes D(1) + A \otimes I_W,$$

$$A = -(1-p)\mu C + \mu(1-p)CE^- + \gamma(NI_{N+1} - C)(E^+ - I) + \beta(I_{N+1} - \tilde{Q})(E^+ - I).$$

Using the so-called mixed product rule for the Kronecker product of matrices, see [42], it is not difficult to verify by the direct substitution that the vector $\mathbf{y}$ that is the solution of the system (10) can be computed in the form

$$\mathbf{y} = \mathbf{x} \otimes \boldsymbol{\theta} \tag{11}$$

where $\boldsymbol{\theta}$ is the invariant probability vector of the stationary distribution of the underlying process of the $BMAP$.

The vector $\mathbf{x}$ is the unique solution to the system

$$\mathbf{x}A = \mathbf{0},\ \mathbf{x}\mathbf{e} = 1.$$

The vector $\mathbf{x}$ is the vector of the stationary distribution of the number of busy servers when the system is overloaded and defines the stationary distribution of the birth-and-death process with the state space $\{0, 1, \ldots, N\}$. The intensities of the birth δ_n and the death ϵ_n in the state n are defined, correspondingly, by Formulas (4) and (5). From the theory of the birth-and-death processes, it is well known that the vector $\mathbf{x}$ in the partitioned form $\mathbf{x} = (x_0, \ldots, x_N)$ is defined by its components given by Formulas (2) and (3). By substituting the vector $\mathbf{y}$ in form (11) into inequality (9), after some algebraic transformations, including the use of mixed product rule, we obtain inequality (1). □

Remark 1. *Ergodicity condition (1) is intuitively clear. The left-hand side of (1) is the rate of customers' arrival to the system. The right-hand side of (1) is the rate of customers' departure from the buffer when the system is overloaded. Indeed, as it was stated above, x_n, $n = \overline{0, N}$, is the probability that n servers are busy at an arbitrary moment when the system is overloaded (correspondingly, $N - n$ servers are on vacation). Under the condition that n servers are busy, the rate of customers departure from the queue is the sum of:*

(i) The rate $np\mu$ of service completion without taking a vacation by the server that finished the service;

(ii) The rate $(N - n)\gamma$ of vacation completion (and starting service) by one of the $N - n$ vacated servers;

(iii) The rate $\beta(1 - q^{N-n})$ of finishing a search of a server having a vacation that is accompanied by an agreement of one of $N - n$ vacated servers to interrupt vacation and start service.

Then (1) reflects the evident condition of stability of the system: when the system is overloaded, customers' arrival rate must be less than the average customers' departure rate from the buffer.

Remark 2. *Ergodicity condition (1) can be used for the preliminary, rough, choice of the system parameters, in particular, the number N of servers. If for the given (or evaluated by some experts)*

intensity λ of arrival, rates of service μ, vacations γ and successful invitations to interrupt vacation, and the fixed value N inequality (1) is not fulfilled, then the number of servers must be increased.

Remark 3. *It is easy to see that the statement of Theorem 2 is valid under assumptions that the following inequalities are fulfilled: $\gamma + \beta > 0$ and $p \neq 1$. In the case of violation of these inequalities, the birth-and-death process with the generator A becomes the death process and $x_0 = 1$ (all servers are permanently vacated) or the birth process and $x_N = 1$ (in the overloaded system all servers are permanently busy). Note that the imposed assumption (from the beginning) that $\gamma > 0$ and $\beta > 0$ can be weakened. One of the rates γ and β can be equal to 0 if another one is positive. If $p = 1$ (the server cannot take a vacation before the queue is exhausted), the servers are no longer self-sustained and the considered queuing system is the usual multi-server queue with vacations. The ergodicity condition is trivial $\lambda < N\mu$.*

5. Computation of the Stationary Distribution of the Markov Chain

Let us assume that the ergodicity condition of the Markov chain ξ_t is fulfilled. Then, the stationary probabilities

$$\pi(i,n,\nu) = \lim_{t\to\infty} P\{i_t = i, n_t = n, \nu_t = \nu\},\ i \geq 0, n = \overline{0,N},\ \nu = \overline{1,W},$$

exists.

Enumerating the states of the Markov chain $\xi_t = \{i_t, n_t, \nu_t\}$ in the direct lexicographic order of the components $\{n_t, \nu_t\}$ and combining the stationary probabilities $\pi(i,n,\nu)$ of the states included into the level i, we obtain the row vectors $\boldsymbol{\pi}_i,\ i \geq 0$.

It is well-known that the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, can be found as the solution to so-called Chapman–Kolmogorov equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)Q = \mathbf{0},\ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)\mathbf{e} = 1. \tag{12}$$

In the case $\alpha = 0$, as it was mentioned above, the generator Q of this Markov chain has an upper-Hessenbergian quasi-Toeplitz structure. The methods for computation of the stationary distribution of such Markov chains are well-known, see, e.g., [39,43]. To make computations for the queuing system under study, we propose the algorithm presented in Section 2.4.4 in [39].

In the case $\alpha > 0$, the solution of (12) encounters tremendous difficulties and we cannot cite any reference in the existing literature where system (12) is solved. Because the generator Q has the upper-Hessenbergian structure, it is possible to write down the recursion by which all the vectors $\boldsymbol{\pi}_i,\ i \geq 1$, can be calculated via the vectors $\boldsymbol{\pi}_j,\ j = \overline{0, i-1}$, and, eventually, via the vector $\boldsymbol{\pi}_0$; however: (i) this recursion is not numerically stable due to multiple sequential implementations of subtraction of sub-stochastic vectors and (ii) what is more important, the impossibility to compute the vector $\boldsymbol{\pi}_0$. Even in the case of $M/G/1$ type Markov chains, the problem of computation of the vector $\boldsymbol{\pi}_0$ is solved based on some additional considerations, e.g., it was offered to use the interpretation of the stationary probability of a state of a discrete-time Markov chain in terms of the average time between the successive visits of the chain to this state. Alternatively, the reasonings of the analyticity of the generating function of the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, in the unit disc of the complex plane are exploited. In the case of the chain with the non-quasi-Toeplitz structure of the generator, both these approaches do not work and no other approach is offered in the literature.

In such a situation, a very popular approach among the researchers in the matrix methods in queuing theory is to make some kind of truncation (rough or soft) of the state space of the chain, for more details see, e.g., [44]. The evident disadvantages of this approach consist of the difficulty of choosing the proper level of truncation and, as a rule, the high dimension of the finite system of equations for the stationary probabilities, which has to be solved numerically.

The analysis of Markov chains having the generator of a structure similar to the generator defined by Theorem 1, which are called in [41] as asymptotically quasi-Toeplitz Markov chains, is very important from the perspective of modeling various systems with customer retrials, see, e.g., [44], systems with impatient customers, systems and tandems of queues with an infinite number of servers; therefore, in [41], a new approach for the computation of the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, is offered. This approach suggests that the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, are computed not via solving the system (12) but by means of a solution of another, specially constructed, system of equations for the vectors $\boldsymbol{\pi}_i,\ i \geq 0$. The procedure of constructing this another system is described in detail in [41]. In two words, this procedure assumes the construction of series of so-called censored Markov chains with respect to the original Markov chain ξ_t with different censoring levels. By means of varying the censoring levels, a new, alternative, system of equations for the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, is obtained.

To solve this system, again, as it was described above with respect to the initial system (12), all vectors $\boldsymbol{\pi}_i,\ i \geq 1$, are recursively expressed via the vector $\boldsymbol{\pi}_0$. This recursion is numerically stable; however, more important is the fact that using the censoring level 0, it is possible to derive the system of equations for the components of the vector $\boldsymbol{\pi}_0$. Under fulfillment of the ergodicity condition, this system supplemented by one more equation derived via the use of normalization condition has a unique solution. The problem of computation of the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, is completely solved.

In [41], more or less comprehensive analysis is implemented for continuous and discrete-time Markov chains, including the account of the possibility of irreducibility and reducibility of certain matrices arising during analysis. Ergodicity conditions and stable numerical algorithms for computation of the stationary distribution of Markov chains are presented.

It is worth mentioning that recently several papers by Japanese authors appeared in which the Markov chains having the upper-Hessenbergian non-quasi-Toeplitz structure of the generator appeared, see, e.g., [45,46]. They also use the idea of derivation of the system of equations for the stationary probabilities or conditional probabilities via the construction of the family of the censored Markov chains; however, in contrast to [41] where the states with large values of a denumerable component are censored, they censor other groups of states. The advantage of [45] over [41] is that no assumption about the form of heterogeneity in the generator is made while in [41] suggestion about the existence of the limiting behavior of the blocks $Q_{i,i+k-1}$ of the generator when i tends to infinity is imposed; however, this advantage turns into a disadvantage because the problem of the derivation of the ergodicity condition is completely solved in [41] and is not touched in [45]. Further, the results from [45] are applicable only if the sub-diagonal blocks $Q_{i,i-1}$ of the generator are non-singular matrices while in [41] they are allowed to be singular. This is very important for the application of the results to the analysis of multi-server queues where these blocks are usually singular.

For the implementation of the numerical experiments described below, we used the algorithms from the recent paper [40] in which the Markov chains having the upper-Hessenbergian non-quasi-Toeplitz structure of the generator are considered. In [40], no assumptions about the form of heterogeneity are made and the presented algorithm for computation of the stationary distribution is much faster than the algorithm based on the results of [41].

6. Performance Indicators

Having computed the row vectors $\boldsymbol{\pi}_i,\ i \geq 0$, and sub-vectors $\boldsymbol{\pi}(i,n)$ defined by $\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \ldots, \boldsymbol{\pi}(i,N))$, we can compute various performance measures of the system. Expressions for some of them are presented below.

The average number of customers in the buffer N_{buffer} can be found as

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

The average number of busy servers N_{server} is computed as

$$N_{server} = \sum_{i=0}^{\infty} \sum_{n=1}^{N} n\boldsymbol{\pi}(i,n)\mathbf{e}.$$

The average number of servers on vacation N_{vacant} is computed as

$$N_{vacant} = N - N_{server}.$$

The average intensity λ_{out} of the output flow of successfully serviced customers can be found as

$$\lambda_{out} = \sum_{i=0}^{\infty} \sum_{n=1}^{N} n\mu\boldsymbol{\pi}(i,n)\mathbf{e} = \mu N_{server}.$$

The loss probability P_{loss} of an arbitrary customer due to impatience is computed as

$$P_{loss} = \frac{1}{\lambda} \sum_{i=1}^{\infty} i\alpha\boldsymbol{\pi}_i \mathbf{e} = \frac{\alpha N_{buffer}}{\lambda} = 1 - \frac{\lambda_{out}}{\lambda}.$$

The probability P_{succ} that the system's search for a vacant server who agrees to interrupt the vacation will be successful is computed as

$$P_{succ} = \frac{\sum\limits_{i=J_2}^{\infty} \sum\limits_{n=0}^{N-1} (1 - q^{N-n})\boldsymbol{\pi}(i,n)\mathbf{e}}{\sum\limits_{i=J_2}^{\infty} \sum\limits_{n=0}^{N-1} \boldsymbol{\pi}(i,n)\mathbf{e}}.$$

The intensity $\sigma_{no\text{-}work}$ of forced vacation of a working server due to empty buffer during service completion epoch can be found as

$$\sigma_{no\text{-}work} = \sum_{n=1}^{N} pn\mu\boldsymbol{\pi}(0,n)\mathbf{e}.$$

The probability $P_{no\text{-}work}$ that after the service completion moment the server would like to start a new service but is forced to take a vacation because the buffer is empty can be found as

$$P_{no\text{-}work} = \frac{\sigma_{no\text{-}work}}{\sum\limits_{i=0}^{\infty} \sum\limits_{n=1}^{N} pn\mu\boldsymbol{\pi}(i,n)\mathbf{e}} = \frac{\sum\limits_{n=1}^{N} n\boldsymbol{\pi}(0,n)\mathbf{e}}{\sum\limits_{i=0}^{\infty} \sum\limits_{n=1}^{N} n\boldsymbol{\pi}(i,n)\mathbf{e}}.$$

The intensity $\sigma_{enough\text{-}servers}$ of the refuses to the servers who would like to start work after vacation completion due to the presence of less than J_1 customers in the buffer can be found as

$$\sigma_{enough\text{-}servers} = \sum_{i=0}^{J_1-1} \sum_{n=0}^{N-1} (N-n)\gamma\boldsymbol{\pi}(i,n)\mathbf{e}.$$

The probability $P_{enough\text{-}servers}$ that the server would like to start work after vacation completion but the system does not give permission to start due to the presence of less than J_1 customers in the buffer can be found as

$$P_{enough\text{-}servers} = \frac{\sigma_{no\text{-}work}}{\sum\limits_{i=0}^{\infty}\sum\limits_{n=0}^{N-1}(N-n)\gamma\boldsymbol{\pi}(i,n)\mathbf{e}} = \frac{\sum\limits_{i=0}^{J_1-1}\sum\limits_{n=0}^{N-1}(N-n)\boldsymbol{\pi}(i,n)\mathbf{e}}{\sum\limits_{i=0}^{\infty}\sum\limits_{n=0}^{N-1}(N-n)\boldsymbol{\pi}(i,n)\mathbf{e}}.$$

7. Numerical Examples

We assume that the arrival flow at the system is the $BMAP$, which is defined by the following matrices

$$D_0 = \begin{pmatrix} -5 & 0 \\ 0 & -1 \end{pmatrix}, D_1 = \begin{pmatrix} 0.2 & 0.02 \\ 0 & 0.2 \end{pmatrix}, D_2 = \begin{pmatrix} 0.08 & 0.1 \\ 0.012 & 0.58 \end{pmatrix},$$

$$D_3 = \begin{pmatrix} 0.58 & 0 \\ 0.002 & 0.2 \end{pmatrix}, D_4 = \begin{pmatrix} 2 & 0.02 \\ 0.002 & 0 \end{pmatrix}, D_5 = \begin{pmatrix} 2 & 0 \\ 0.002 & 0.002 \end{pmatrix},$$

and has the following characteristics:

The average arrival rate of customers $\lambda = 4.11215$;

The average arrival rate of batches $\lambda_b = 1.4557$;

The coefficient of correlation of successive inter-arrival times $c_{cor} = 0.187215$;

The squared coefficient of variation of inter-arrival times $c_{var} = 1.64605$.

The service intensity μ is assumed to be equal to 0.3. The intensity of each customer's impatience is $\alpha = 0.01$.

The parameter γ of the exponential distribution of the vacation time is equal to 0.1. The intensity of each customer's impatience is $\alpha = 0.01$. The probability p that the server starts a new service after the service completion, if the buffer is not idle, is equal to 0.5. The parameter β of the exponential distribution of the search time is $\beta = 0.5$. The probability that each vacated server declines the offer to interrupt the vacation and start to work is $q = 0.7$.

The aim of the numerical example is to define the control parameters N, J_1 and J_2, which maximize the following economical criterion:

$$E(N, J_1, J_2) = a\lambda_{out} - b\lambda P_{loss} - c\sigma_{no\text{-}work} - d\sigma_{enough\text{-}servers}.$$

Here, a is a profit gained by the system from the service of one customer, b is a charge for the loss of one customer, c is a charge paid by the system when the server wants to start a new service after the service completion, but the buffer is empty and d is a charge paid by the system when the vacant server wants to start work, but the number of customers in the buffer is less than J_1.

In the numerical example, we fix the following values of costs coefficients:

$$a = 2,\ b = 5,\ c = 1,\ d = 0.5.$$

Let us assume that the number of servers N can take the value from the range [5, 200], the parameter J_1 can vary from 1 to 20, and the parameter J_2 can vary from J_1 to 25.

Computations implemented under the fixed above values of parameters of the system and values N, J_1, J_2 show that the maximal value E^* of the economical criterion $E(N, J_1, J_2)$ is $E^* = E(65, 3, 11) = 4.38728$. Thus, to maximize the profit of the system it is required to have $N^* = 65$ servers, allow for the vacant server to start work when at least $J_1^* = 3$ customers stay in the buffer and start the search of a server who wants to work among vacant servers when the number of customers in the buffer is more or equal to $J_2^* = 11$.

Note, that if the manager of the system does not perform any control, i.e., the parameter $J_1 = 1$ (each vacant server can start work if the buffer is not empty) and the intensity $\beta = 0$ (the system never searches servers who wants to start work), the optimal value of the cost criterion is $E^* = 4.10935$ and achieved for $N = 68$.

Because it is not possible to present 4D figures, let us fix the value of the parameter $J_2 = J_2^* = 11$, and vary the number of servers N over the interval [5, 200] and the parameter J_1 over the interval $[1, J_2]$.

Figure 2 illustrates the dependence of the average number of customers in the buffer N_{buffer} on the number of servers N and the parameter J_1.

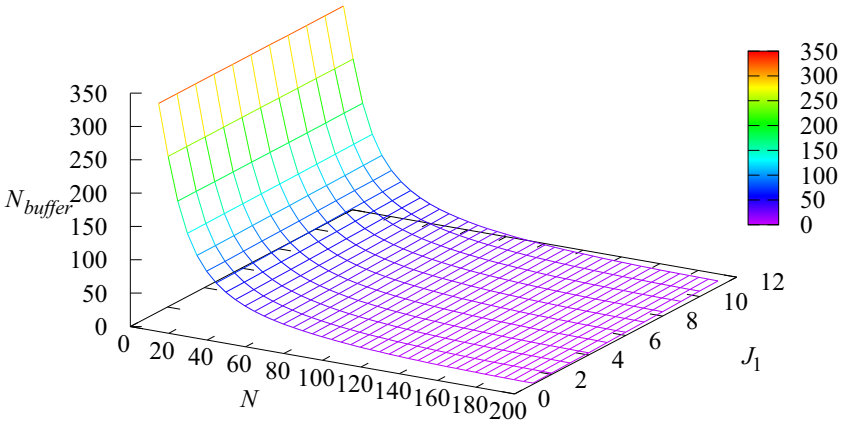

Figure 2. Dependence of the average number of customers in the buffer N_{buffer} on N and J_1.

As it is seen from Figure 2, the average number of customers in the buffer N_{buffer} decreases with the increase in the number of servers N and increases with the increase in the parameter J_1.

Figure 3 illustrates the dependence of the average number of busy servers N_{server} on the number of servers N and the parameter J_1.

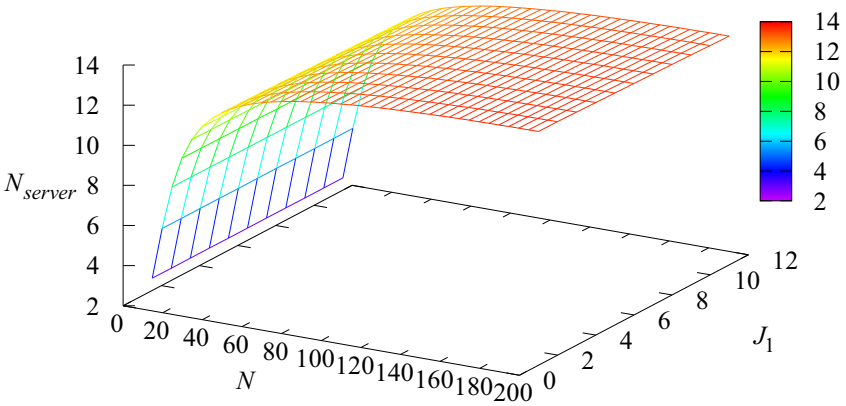

Figure 3. Dependence of the average number of busy servers N_{server} on N and J_1.

As it is seen from Figure 3, the average number of busy servers N_{server} decreases with the increase in the parameter J_1 and increases with the increase in the number of servers N.

Figure 4 illustrates the dependence of the loss probability P_{loss} on the number of servers N and the parameter J_1.

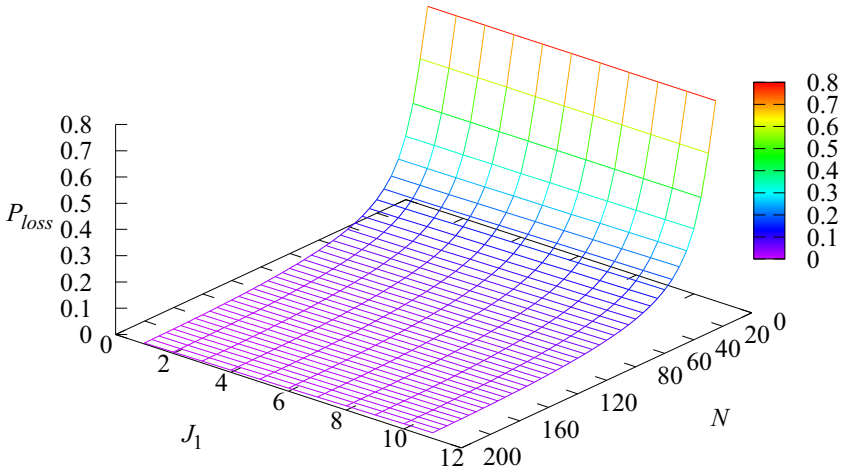

Figure 4. Dependence of the loss probability P_{loss} on N and J_1.

As it is seen from Figure 4, the loss probability P_{loss} increases with the increase in the parameter J_1 and decreases with the increase in the number of servers N.

The dependence of the probability $P_{no\text{-}work}$ that after the service completion moment the server would like to start a new service but is forced to take a vacation because the buffer is empty on the number of servers N and the parameter J_1 is illustrated in Figure 5.

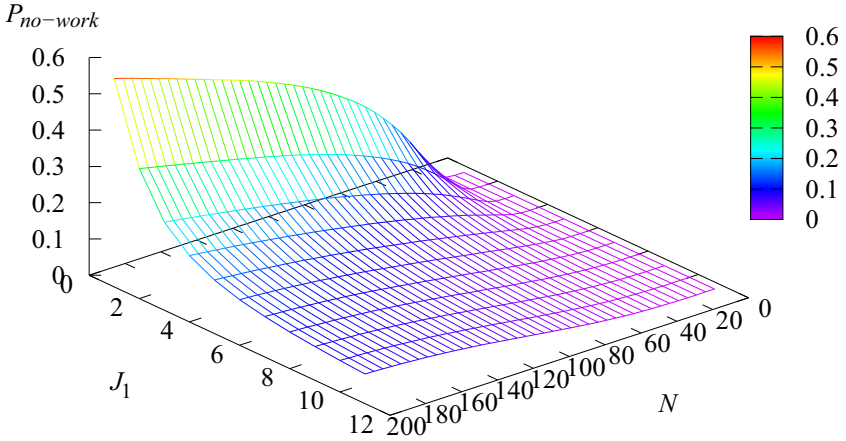

Figure 5. Dependence of the probability $P_{no\text{-}work}$ on N and J_1.

One can see from Figure 5 that the probability $P_{no\text{-}work}$ grows with the increase in the number of customers and decreases when the parameter J_1 increases.

Figure 6 illustrates the dependence of the probability $P_{enough\text{-}servers}$ that the server would like to start work after vacation completion but the system does not give permission to start due to the presence of less than J_1 customers in the buffer on the number of servers N and the parameter J_1.

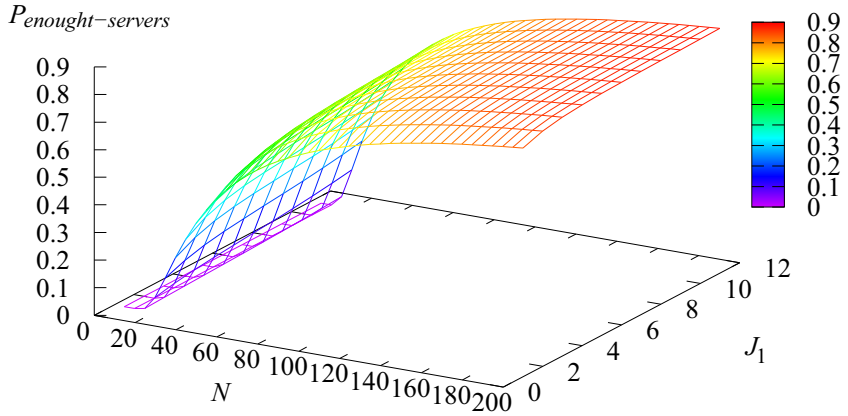

Figure 6. Dependence of the probability $P_{enough\text{-}servers}$ on N and J_1.

As it is seen from Figure 6, the probability $P_{enough\text{-}servers}$ increases with the increase in the parameter J_1 and the number of servers N.

The dependence of the values of the cost criterion E on the number of servers N and the parameter J_1 is illustrated in Figure 7.

Some of the dependencies presented in Figures 2–7 are quite *qualitatively* clear. The values of Figures 2–7 characterize these dependencies quantitatively, which allows to solve various optimization problems.

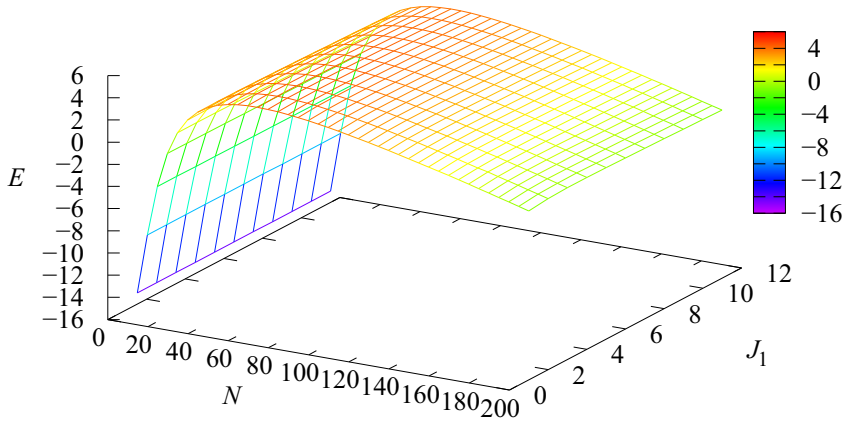

Figure 7. Dependence of the probability $P_{enough\text{-}servers}$ on N and J_1.

8. Conclusions

In this paper, we introduce into consideration a new kind of multi-server queuing systems with servers vacation. A novel feature of the model is a high degree of independence of servers of the decision-maker of the system. Any server from the finite pool of servers can stop operation after service completion of another customer and go for vacation. The manager of the system can implicitly influence performance of the system via the proper choice of the thresholds defining the providing to a server an opportunity to start work after vacation completion. Further, the manager can use the right to try to interrupt the ongoing vacation of a server if congestion occurs. The arrival flow is assumed to be defined by the *BMAP* that allows fitting modern flows. Such a type of model can have potential applications for the investigation of real systems with low centralization of operation of the system and the possibility of flexible choice of a working schedule by the servers, e.g., some modern systems with workers that are the freelancers that work at their free time when they wish, e.g., transportation systems, in particular, taxi drivers.

The model deserves to be generalized in many directions including consideration of the phase-type distribution of service or search times (which allows to more exactly fit the statistics about the values of these times, comparing the exponential distribution suggested

in this paper), more involved strategies of control by beginning/finishing vacations, presence of a certain pool of permanent servers that completely obey the manager, account of retrial phenomenon (see [44,47,48]), random fluctuation of the system parameters, etc.

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

Funding: The publication was prepared with the support of 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. Mitrani, I.L.; Avi-Itzhak, B. A many-server queue with server interruptions. *Oper. Res.* **1968**, *163*, 628–638. [CrossRef]
2. Neuts, M.F.; Lucantoni, D.M. A Markovian queue with N servers subject to breakdowns and repairs. *Manag. Sci.* **1979**, *25*, 849–861. [CrossRef]
3. Krishnamoorthy, A.; Pramod, P.K.; Chakravarthy, S.R. Queues with interruptions: A survey. *Top* **2014**, *22*, 290–320. [CrossRef]
4. Chakravarthy, S.R.; Kulshrestha, R. A queueing model with server breakdowns, repairs, vacations, and backup server. *Oper. Res. Perspect.* **2020**, *7*, 100131. [CrossRef]
5. Kim, C.S.; Klimenok, V.I.; Dudin, A.N. Analysis of unreliable $BMAP/PH/N$ type queue with Markovian flow of breakdowns. *Appl. Math. Comput.* **2017**, *314*, 154–172. [CrossRef]
6. Chakravarthy, S.R.; Ozkar, S.; Shruti, S. Analysis of $M/M/c$ Retrial Queue with Thresholds, PH Distribution of Retrial Times and Unreliable Servers. *J. Appl. Math. Inform.* **2021**, *39*, 173–196.
7. Yang, X.; Alfa, A.F. A class of multi-server queueing systems with server failures. *Comput. Ind. Eng.* **2009**, *56*, 33–43. [CrossRef]
8. Klimenok, V.I.; Orlovsky, D.S.; Kim, C.S. The $BMAP/PH/N$ retrial queue with Markovian flow of breakdowns. *Eur. J. Oper. Res.* **2008**, *189*, 1057–1072.
9. Dudin, A.; Dudin, S. Analysis of a priority queue with phase-type service and failures. *Int. J. Stoch. Anal.* **2016**, *2016*, 9152701. [CrossRef]
10. Kim, C.; Dudin, A.; Dudin, S.; Dudina, O. Analysis of an $MMAP/PH_1, PH_2/N/\infty$ queueing system operating in a random environment. *Int. J. Appl. Math. Comput. Sci.* **2014**, *24*, 485–501. [CrossRef]
11. Doshi, B.T. Queueing systems with vacations—A survey. *Queueing Syst.* **1986**, *1*, 29–66. [CrossRef]
12. Tian, N.Z.; Zhang, G. *Vacation Queueing Models-Theory and Applications*; Springer: Heidelberg, Germany, 2006.
13. Chao, X.; Zhao, Y.Q. Analysis of multi-server queues with station and server vacations. *Eur. J. Oper. Res.* **1998**, *110*, 392–406. [CrossRef]
14. Kim, B.; Kim, J.; Bueker, O. Non-preemptive priority $M/M/m$ queue with servers' vacations. *Comput. Ind. Eng.* **2021**, *160*, 107390. [CrossRef]
15. Chakravarthy, S.R. A Comparative Study of Vacation Models Under Various Vacation Policies: A Simulation Approach. In *Mathematical Modeling and Computation of Real-Time Problems*; CRC Press: Boca Raton, FL, USA, 2021; pp. 3–20.
16. Kim, C.S.; Dudin, A.; Dudina, O.; Klimenok, V. Analysis of queueing system with non-preemptive time limited service and impatient customers. *Methodol. Comput. Appl. Probab.* **2020**, *22*, 401–432. [CrossRef]
17. Takagi, H. *Queueing Analysis: A Foundation of Performance Evaluation*; North-Holland: Amsterdam, The Netherlands, 1991.
18. Dudin, A.; Kim, C.S.; Dudin, S.; Dudina, O. Priority Retrial Queueing Model Operating in Random Environment with Varying Number and Reservation of Servers. *Appl. Math. Comput.* **2015**, *269*, 674–690. [CrossRef]
19. Kim, C.S.; Dudin, S. Analysis of Type Queueing System Operating in Random Environment. *J. Korean Inst. Ind. Eng.* **2016**, *42*, 30–37.
20. Gnedenko, B.V.; Kovalenko, I.N. *Introduction in Queueing Theory*; KVIRTU: Kiev, Ukraine, 1963. (In Russian)
21. Bin, S.; Dudin, S.A.; Dudina, O.S.; Dudin, A.N. A Customer Service Model in an Adaptive-Modulation Mobile Communication Cell with Allowance for Random Environment. *Autom. Remote Control* **2021**, *82*, 812–826. [CrossRef]
22. Kim, C.S.; Klimenok, V.; Mushko, V.; Dudin, A. The $BMAP/PH/N$ retrial queueing system operating in Markovian random environment. *Comput. Oper. Res.* **2010**, *37*, 1228–1237. [CrossRef]
23. Yang, G.; Yao, L.G.; Ouyang, Z.S. The $MAP/PH/N$ retrial queue in a random environment. *Acta Math. Appl. Sin. Engl. Ser.* **2013**, *29*, 725–738. [CrossRef]

24. Lui, J.C.S.; Golubchik, L. Stochastic complement analysis of multi-server threshold queues with hysteresis. *Perform. Eval.* **1999**, *35*, 19–48. [CrossRef]
25. Li, H.; Yang, T. Queues with a variable number of servers. *Eur. J. Oper. Res.* **2000**, *124*, 615–628. [CrossRef]
26. Kim, C.S.; Dudin, A.; Dudin, S.; Dudina, O. Hysteresis Control by the Number of Active Servers in Queueing System with Priority Service. *Perform. Eval.* **2016**, *101*, 20–33. [CrossRef]
27. Mitrani, I. Trading power consumption against performance by reserving blocks of servers. In *Computer Performance Engineering*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 1–15.
28. Green, L. A queueing system in which customers require a random number of servers. *Oper. Res.* **1980**, *28*, 1335–1346. [CrossRef]
29. Kim, C.S.; Lee, M.H.; Dudin, A.; Klimenok, V. Multi-server queueing systems with cooperation of the servers. *Ann. Oper. Res.* **2008**, *162*, 57–68. [CrossRef]
30. Dudin, A.N.; Sun, B. Unreliable Multi-server System with Controllable Broadcasting Service. *Autom. Remote Control* **2009**, *70*, 2073–2084. [CrossRef]
31. Krishnamoorthy, A.; Manikandan, R.; Shajin, D. Analysis of a multiserver queueing-inventory system. *Adv. Oper. Res.* **2015**, *2015*, 747328. [CrossRef]
32. Krishnamoorthy, A.; Shajin, D.; Lakshmy, B. On a queueing-inventory with reservation, cancellation, common life time and retrial. *Ann. Oper. Res.* **2016**, *247*, 365–389. [CrossRef]
33. Gelenbe, E. Synchronising energy harvesting and data packets in a wireless sensor. *Energies* **2015**, *8*, 356–369. [CrossRef]
34. Dudin, A.N.; Lee, M.H.; Dudin, S.A. Optimization of Service Strategy in Queueing System with Energy Harvesting and Customers Impatience. *Int. J. Appl. Math. Comput. Sci.* **2016**, *26*, 367–378. [CrossRef]
35. Baek, J.H.; Dudina, O.; Kim, C.S. Queueing System with Heterogeneous Impatient Customers and Consumable Additional Items. *Int. J. Appl. Math. Comput. Sci.* **2017**, *27*, 367–384. [CrossRef]
36. Sun, B.; Dudin, A.; Dudin, S. Queueing system with impatient customers, visible queue and replenishable inventory. *Appl. Comput. Math.* **2018**, *17*, 161–174.
37. Chakravarthy, S.R. The batch Markovian arrival process: A review and future work. In *Advances in Probability Theory and Stochastic Processes*; Krishnamoorthy, A., Raju, N., Ramaswami, V., Eds.; Notable Publications Inc.: Princeton, NJ, USA, 2001; pp. 21–29.
38. Lucantoni, D. New results on the single server queue with a batch Markovian arrival process. *Commun. Statist.-Stoch. Model.* **1991**, *7*, 1–46. [CrossRef]
39. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer Nature: Heidelberg, Germany, 2019.
40. Dudin, S.; Dudin, A.; Kostyukova, O.; Dudina, O. Analysis of single-server retrial queueing model with a finite buffer, batch arrivals, and unreliable service. *J. Comput. Appl. Math.* **2020**, *366*.
41. 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]
42. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Horwood, E., Ed.; Courier Dover Publications: Cichester, UK, 1981.
43. Neuts, M.F. *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*; Marcel Dekker: New York, NY, USA, 1989.
44. Falin, G.; Templeton, J.G.C. *Retrial Queues*; CRC Press: Boca Raton, FL, USA, 1997.
45. Takine, T. Analysis and computation of the stationary distribution in a special class of Markov chains of level-dependent M/G/1-type and its application to $BMAP/M/\infty$ and $BMAP/M/c+M$ queues. *Queueing Syst.* **2016**, *84*, 49–77. [CrossRef]
46. Masuyama, H. A sequential update algorithm for computing the stationary distribution vector in upper block-Hessenberg Markov chains. *Queueing Syst.* **2019**, *92*, 173–200. [CrossRef]
47. Breuer, L.; Dudin, A.N.; Klimenok, V.I. A retrial $BMAP/PN/N system$. *Queueing Syst.* **2002**, *40*, 433–457. [CrossRef]
48. Breuer, L.; Klimenok, V.; Birukov, A.; Dudin, A.; Krieger, U.R. Modeling the access to a wireless network at hot spots. *Eur. Trans. Telecommun.* **2005**, *16*, 309–316. [CrossRef]