mathematics

MDPI

Article

Analysis of $BMAP/PH/N$-Type Queueing System with Flexible Retrials Admission Control

Sergei A. Dudin [1,*], Olga S. Dudina [1], Azam A. Imomov [2] and Dmitry Y. Kopats [3]

[1] Department of Applied Mathematics and Computer Science, Belarusian State University, 4 Nezavisimosti Ave., 220030 Minsk, Belarus; dudina@bsu.by

[2] Physics and Mathematical Faculty, Karshi State University, 17 Kuchabag Str., Karshi City 180100, Uzbekistan; imomov_azam@mail.ru

[3] Faculty of Mathematics and Computer Science, Yanka Kupala State University of Grodno, 22 Orzheshko Str., 230027 Grodno, Belarus; dk80395@mail.ru

[*] Correspondence: dudins@bsu.by

Abstract: This research examines a multi-server retrial queueing system with a batch Markov arrival process and a phase-type service time distribution. The system's distinguishing feature is its ability to control the admission of retrial customers. An attempt by a customer to retry is successful only if the number of busy servers does not exceed certain threshold values, which may depend on the state of the fundamental process of the primary customer's arrival. Impatient retrying customers may abandon the system without obtaining service. A group of primary customers that arrives while the number of available servers is fewer than the group size is either entirely rejected or occupies all available servers, while the remainder of the group transitions to the orbit. The system's behavior, under a defined set of thresholds, is characterized by a multidimensional Markov chain classified as asymptotically quasi-Toeplitz. This enables the acquisition of the ergodicity condition and the computation of the steady-state distribution of the Markov chain and the system's performance measures. The presented numerical examples demonstrate the impact of threshold value variation. An example of solving an optimization problem is presented. The importance of the account of the batch arrivals is shown.

Keywords: batch Markov arrival process; phase-type service time distribution; impatience; retrial admission control; asymptotically quasi-Toeplitz Markov chain

MSC: 60K25; 60K30; 68M20; 90B22

check for updates

Academic Editor: Lijun Bo

Received: 17 March 2025
Revised: 24 April 2025
Accepted: 25 April 2025
Published: 27 April 2025

Citation: Dudin, S.A.; Dudina, O.S.; Imomov, A.A.; Kopats, D.Y. Analysis of $BMAP/PH/N$-Type Queueing System with Flexible Retrials Admission Control. *Mathematics* **2025**, *13*, 1434. https://doi.org/10.3390/math13091434

1. Introduction

Retrial queues are characterized by the feature that there is no place for waiting by arriving customers in the case of the unavailability of service facilities at the arrival moment. However, a rejected customer may relocate to a virtual location known as "orbit" and attempt to access the service at random intervals. For instance, the monograph [1] contains examples of the theory of retrial queues applied to analyze real-world systems. As an early work in the theory of retrial queues, [2] has to be mentioned. Surveys of the research on retrial queues are given in [3–10]. Single-server retrial queues are investigated (see the excellent monograph [4]) to a much better extent than queues with many servers due to the essentially higher dimension of the stochastic processes describing the multi-server retrial queues. In this paper, we consider a multi-server retrial queue.

The predominant body of study in queueing theory, particularly concerning retrial queues, focuses on systems with a stationary Poisson arrival process defined by a constant customer arrival rate. It is already well recognized that an arrival rate can fluctuate in many real-world systems. As a good model of arrivals with varying instantaneous arrival rates, the Markov arrival process (MAP) and its generalization, such as the batch Markov arrival process ($BMAP$), are introduced in, e.g., [11,12]. For more information about the $BMAP$, its properties, usefulness for modeling real-world flows, and literature surveys, see [13–17]. The modeling of real-world flows by the $BMAP$ allows one to catch the inherent features of such flows as the fluctuating arrival rate, dependence between successive inter-arrival times, and possible high variance of these times. The issue of matching real-world flows with the $BMAP$ has been thoroughly examined in the current literature; refer to, for instance, [17–19] and the citations therein.

The initial investigation of a multi-server retrial queue characterized by a $BMAP$ and phase-type (PH) service time distribution is presented in [20]. Phase-type distribution is an essential generalization of the most popular distribution among researchers, exponential distribution. In contrast to this distribution, which allows for the fitting of only the expectations of service times, PH distribution allows for the fitting of their variance and higher moments. For more information about PH distribution, see [21,22]. An approximation of any arbitrary distribution of a non-negative random variable can be accomplished through the utilization of the PH distribution (see [23]). The condition for the ergodicity of the $BMAP/PH/N$ retrial queue is established in paper [20]. A very important problem of the computation of this distribution is not touched upon there due to its complexity. This problem was later solved in [24]. In [24], sufficient ergodicity conditions and algorithms for calculating the stationary distribution of system states are given for the cases of a constant and infinitely increasing total retrial rates from the orbit.

The power of queueing theory consists in its ability to provide a tool not only for the probabilistic analysis of real-world systems under the fixed parameters of arrival, service, retrial, etc., processes but also for the elaboration of proposals on how these parameters might be modified to enhance the operation of the systems. Besides the traditional static optimization problems, e.g., finding the minimal number of servers sufficient to guarantee the fixed requirements of the quality of service indicators, which were solved in the beginning of queueing theory by its founder A.K. Erlang, currently, the problems of dynamic optimization are also actual. The swift advancement of electronic and communication technologies now enables the dynamic monitoring of queueing systems (e.g., the count of occupied servers, queue length, or the number of customers in the orbit) and facilitates immediate responses to unfavorable state changes. For instance, if the customer volume in the system escalates excessively, one may augment the number of servers delivering service, enhance the service rates of existing servers, or, alternatively, reduce the arrival rate, etc.

The standard retrial queueing models assume equal opportunities to occupy the idle servers for the primary and retrying customers. However, in various potential real-world applications, it is likely reasonable differentiate these opportunities different due to economic reasons. An analogous situation takes place in modeling so-called cognitive radio systems in which all types of customers may require the same number of system resources, but one of the types, the primary customers, have priority access as the owners of the channels. Another type, so-called cognitive customers, have opportunistic access and may obtain access only if a sufficient number of existing channels are idle.

By analogy, it seems to be reasonable to give some priority access to the primary customers in multi-server retrial queues over the retrying customers. The motivation for this is twofold. The first reason is that the main characteristic of quality of service in many systems is, namely, the probability of immediate access to service of an arbitrary arriving customer. Providing priority to the primary customers may increase this probability. The second economical reason is the following. If the primary customer is rejected, this customer may decide to abandon service in the system and permanently leave it without bringing any revenue to the system owner. If a customer decided to become a retrial customer, maybe he/she has a serious need to receive service in this system and will not leave the system without service and bringing the profit.

Priority access for the primary customers can be managed via the reservation of a few servers exclusively for the access of the primary customers when the number of idle servers is small, similar to the existing schemes of channel reservation in cognitive radio systems. The reservation makes sense even in the case when the rate of customer arrival is constant.

The use of the $BMAP$ as the model of a real-world arrival flow can ensure a more effective implementation of control based not solely on the present quantity of customers in service or within the system but also on the state of the fundamental process of the $BMAP$. Specifically, in the context of multi-server retrial queues, it is prudent to regulate the retrial rate based on the number of occupied servers and the state of the arrival process. An interesting specific instance of the $BMAP$ is the Markov modulated Poisson process ($MMPP$). In the $MMPP$, during some exponentially distributed time interval, arrivals occur like, in the stationary Poisson process with some fixed rate. When this interval finishes, the rate admits another value. In the simplest case, we can speak about the intervals with high and low arrival rates. If primary customers can be lost when all servers are busy at their arrival moment, it makes sense to restrict the speed of retrials during the periods when the quantity of occupied servers is large and allow for a higher retrial rate when the number of busy servers is relatively smaller and a quick occupation of idle servers by the customers from orbit is desirable. The correctness of these intuitive considerations is confirmed in [25], in which the $MAP/M/N$ retrial queue is studied. The threshold-type strategy is employed to implement the admission of retrying customers for service. A retrying customer is acceptable only if the count of occupied servers does not surpass the threshold associated with the present state of the fundamental process of the MAP. Alternatively, the customer reverts to the orbit. The system's behavior, given a fixed set of thresholds, is characterized by a three-dimensional Markov chain whose stationary characteristics are analyzed in [25]. The possibility of optimally choosing the thresholds is numerically demonstrated.

We classify [25] as an opening of a new direction in dynamic control by queueing systems, including retrial queues, depending not only on the state of the orbit or buffer and servers but also on the state of the fundamental process of arrivals. Traditionally, control is executed solely based on the state of the orbit or the buffer and servers, as fluctuations in the arrival process are seldom considered. While these fluctuations are quite typical for many real-world systems, fluctuations in arrival rate can be caused by time intervals during the day or night, as well as the day of the week, weather conditions, and other random events and factors. Our paper investigates this new direction and essentially extends the results of [25].

Two assumptions, which are quite restrictive from the point of view of potential applications, were imposed in the model analyzed in [25]: service times have exponential distribution and customers can arrive only one by one. In this paper, we essentially relax both these assumptions. We suppose a much more general PH distribution of service times, which allows for a better fit for the real distribution of the service time, or at least the variance of this time. Customers may arrive both singly and in random-sized batches, with their arrival characterized by the $BMAP$. Batch arrival is characteristic of numerous real-world systems, e.g., telecommunication networks, where the arriving message represents a group of packets, or emergency services, where a group arrival of patients may occur as the result of an accident. It is clear that the possibility of batch arrivals of customers and their rejection when the number of available servers is insufficient for service of the arrived batch makes the reservation of servers (and access denial to retrials) for the primary customer service even more important than in the model considered in [25]. The numerical experiments, including those provided in this study, conclusively demonstrate that the ignorance of batch arrivals (by replacing the $BMAP$ by the MAP with the same mean arrival rate) leads to too optimistic an evaluation of the performance of the system and an incorrect estimation of the amount of the resource required to provide the required quality of service level.

Thus, in this paper, we examine the $BMAP/PH/N$ retrial queue. Analysis of this system is essentially more complicated than the analysis implemented in [25] due to two reasons.

The first one is that the Markov chain describing the dynamics of the system has a higher dimension of a state space. In [25], this chain has three components: the number of customers in orbit, the number of customers in service, and the fundamental process of the MAP. Here, several additional components, which characterize the service process in busy servers, are required, and the cardinality of the state space of these components can be pretty large.

The second reason is that customers may arrive in groups, resulting in the generator of the chain lacking a block tridiagonal structure and the chain not being classified as level-dependent quasi-birth-and-death processes (see [21]) due to the potential for retrials and the impatience of orbiting customers.

This paper's content is succinctly outlined as follows. The mathematical model delineating the system's behavior is presented in Section 2. Section 3 is devoted to the description of a Markov chain that indicates the dynamics of the examined system. This section presents and elucidates the explicit form of the generator in detail. The affiliation of this chain with the category of asymptotically quasi-Toeplitz Markov chains is confirmed and the ergodicity of this chain for any configuration of the system parameters is demonstrated. The issue of calculating the stationary distribution of the chain is briefly addressed. Section 4 presents formulas for calculating the key performance characteristics of the system. Section 5 delineates the outcomes of two numerical experiments. In the first experiment, the system with the $BMAP$, the fundamental process of which has two states and, therefore, control by retrials admission is defined by two thresholds, is considered, and the dependence of several performance characteristics of the system on these thresholds is numerically highlighted. An example of solving an optimization problem is presented. In the second experiment, the following question is answered: whether or not the generalization of the results from [25] to the case of batch arrivals is essential. The answer is that the batch arrivals do matter. Section 6 summarizes this paper and lists possible generalizations of the considered model.

2. Mathematical Model Description

We examine a retrial multi-server queuing system, the configuration of which is illustrated in Figure 1.

The $BMAP$ flow of primary customers enters the system. This arrival process is characterized by the fundamental process ν_t, $t \geq 0$, which is an irreducible continuous-time Markov chain with a finite state space $\{1, 2, \dots, V\}$, and the matrices D_k, $k = \overline{1, K}$, which represent the jump intensities of the process ν_t associated with the arrival of a batch of k customers. Here, K is the maximum feasible batch size. The notation $k = \overline{1, K}$ indicates that the parameter k can take on values from the set $\{1, 2, \dots, K\}$. The matrix $D(1) = \sum\limits_{k=0}^{K} D_k$ denotes the generator of the process ν_t.

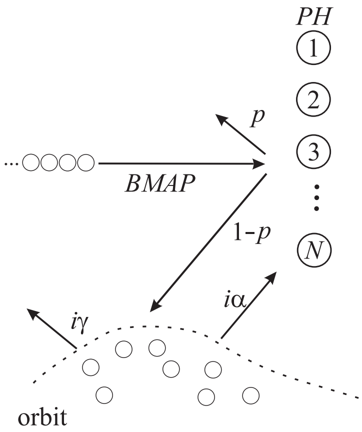

Figure 1. Structure of the system under study.

The average arrival rate (fundamental rate) λ of the $BMAP$ is determined by the equation $\lambda = \boldsymbol{\theta} \sum\limits_{k=1}^{K} kD_k\mathbf{e}$, where $\boldsymbol{\theta}$ represents the invariant vector of the stationary distribution of the process ν_t. The vector is determined as the single solution to the system $\boldsymbol{\theta}D(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Here and below, $\mathbf{e}$ denotes a column vector comprised of ones, while $\mathbf{0}$ represents a zero row vector. If the dimension of the vectors $\mathbf{e}$ and $\mathbf{0}$ is not clear from the context, it may be indicated by the subscript, e.g., $\mathbf{0}_V$ means a zero vector of size V.

Formulas for calculating the coefficients of variation and correlation of inter-arrival times are available in sources such as [12–14,16].

The system does not have a buffer and has N identical independent servers that provide service to customers. The service time for any customer by any server has a phase-type distribution characterized by the irreducible representation $(\boldsymbol{\beta}, S)$. The distribution is defined by the Markov chain m_t, $t \geq 0$, which has a state space comprising transient states $\{1, 2, \dots, M\}$ and an absorbing state $M + 1$. The duration exhibiting this distribution can be regarded as the walking time of a particle within a network including M nodes. The stochastic row vector $\boldsymbol{\beta}$ of dimension M defines the starting position of the particle in the network at the walking beginning instant. The square matrix S of dimension M is the subgenerator that delineates the transition rates of the chain m_t within the set $\{1, 2, \dots, M\}$. The transition to the absorbing state signifies the conclusion of the walking process. The transition rates to the absorbing states are represented by the column vector $\mathbf{S}_0 = -S\mathbf{e}$. The average service time of a customer is defined as $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$.

Customer processing is described as follows. If a group of customers arrives and discovers the necessary number of idle servers, it begins processing on these servers. If the quantity of idle servers is fewer than the size of the incoming batch, then this batch is rejected and leaves the system permanently with probability p or, with the complementary probability, part of the batch occupies all idle servers, if any, while the remainder of the batch joins the orbit. The orbit is a virtual place from which customers individually and independently of each other make repeated attempts (retrials) to enter the service. The orbit capacity is assumed to be infinite. Inter-retrial times have an exponential distribution with the parameter α, $\alpha > 0$. When i customers stay in the orbit, the total retrial rate is $i\alpha$.

In the standard retrial queues, if a retrial is successful, the customer leaves the orbit and starts service if there is at least one idle server. We make the following assumption about a more strict restriction on customers' access from the orbit.

We fix the set of integer parameters (thresholds) R_ν, assuming that $0 \leq R_\nu < N$, $\nu = \overline{1, V}$. If, during the retrial epoch, the status of the $BMAP$ arrival process is ν and the number of occupied servers does not exceed the threshold R_ν, then this customer starts service on one of the idle servers. Otherwise, the customer remains in the orbit awaiting service.

Customers in the orbit may exhibit impatience and depart from the orbit and system without service independently after an exponentially distributed time characterized by the parameter γ, $\gamma > 0$.

The further analysis aims to find the steady-state distribution of the considered system under any fixed set of the thresholds $\{R_1, R_2, \dots, R_V\}$ and examine how these thresholds affect the system's performance characteristics.

3. The Process of System States and Its Analysis

The system's behavior can be characterized by the regular irreducible continuous-time Markov chain

$$\xi_t = \{i_t, \nu_t, n_t, m_t^{(1)}, \dots, m_t^{(M)}\}, \; t \geq 0,$$

where, at time instant t,

- i_t denotes the quantity of orbital customers, $i_t \geq 0$;
- ν_t denotes the state of the fundamental process of the $BMAP$, $\nu_t = \overline{1, V}$;
- n_t denotes the number of occupied servers, $n_t = \overline{0, N}$;
- $m_t^{(l)}$ denotes the quantity of servers in the l-th phase of service, $m_t^{(l)} = \overline{0, n_t}$, $l = \overline{1, M}$, $\sum\limits_{l=1}^{M} m_t^{(l)} = n_t$.

It is important to acknowledge that the assumption about the PH distribution of service time rather than its specific instance, the exponential distribution, which is assumed in [25], essentially complicates the form of the Markov chain that should be analyzed. In [25], three components, $\{i_t, \nu_t, n_t\}$, completely characterize the dynamics of the queueing system. In the case of the PH distribution of the service time, this three-dimensional process is not Markovian. To receive the Markovian process, it is necessary to supplement these components with the auxiliary process with components describing the state of the service process in busy servers.

Two popular ways for receiving the Markov process are as follows. The initial method is referred to as $TPFS$ (track phase for server); see [26], which presumes the continuous re-enumeration of occupied servers according to their use and the monitoring of the service phase for each occupied server.

Consequently, when the number of customers receiving service equals n, the auxiliary process must consist of the components $\boldsymbol{\zeta}_t = (\zeta_t^{(1)}, \zeta_t^{(2)}, \dots, \zeta_t^{(n)})$, $t \geq 0$, where $\zeta_t^{(r)}$ represents the current phase of service for the rth active server, with $\zeta_t^{(r)} = \overline{1, M}$, for $r = \overline{1, n}$. The cardinality of the state space of the process $\boldsymbol{\zeta}_t$ when the number of customers receiving service equals n is equal to M^n, $n = \overline{1, N}$.

The second method, referred to as $CSFP$ (count server for phase) in the formulation of the service process, entails monitoring the quantity of servers delivering service throughout various stages. Building upon the methodology established in [27], we ascertain that when the quantity of customers receiving service is n, the auxiliary process is delineated as $\mathbf{m}_t = \{m_t^{(1)}, \dots, m_t^{(M)}\}$, $t \geq 0$, where the component $m_t^{(k)}$ represents the number of servers in the kth service phase, with $k = \overline{1, M}$. The cardinality T_n of the state space of the process $\mathbf{m}_t$ when the number of customers receiving service equals n is given by

$$T_n = \binom{n + M - 1}{M - 1} = \frac{(n + M - 1)!}{n!(M - 1)!}, \quad n = \overline{1, N}.$$

By default, we suppose that $T_0 = 1$.

As follows from the definition of the process ξ_t given above, we chose the second way to construct the Markov process. This choice can be explained by the fact that although the second way is characterized by the less transparent formulas for transition intensities, the cardinality of the state space of the process $\mathbf{m}_t$ can be drastically smaller, especially in the case of small values of M, than the cardinality of the state space of the process $\boldsymbol{\zeta}_t$; see, e.g., [26]. For example, if $M = 2$ and $n = 10$, then the cardinality of the state space of the process $\boldsymbol{\zeta}_t$ is equal to $2^{10} = 1024$. The cardinality of the state space of the process $\mathbf{m}_t$ is $n + 1 = 11$.

The cardinality of the state space of the process describing the service of customers is critically significant during the computer implementation phase of the devised technique for calculating the stationary distribution of the Markov chain, particularly when dealing with a substantial number of servers N. This motivated our choice of the $CSFP$ approach in this study.

We list the states of the Markov chain $\xi_t,\ t \geq 0$ in direct lexicographic order based on the components $\{i_t, \nu_t, n_t\}$ and in reverse lexicographic order based on the components $\{m_t^{(1)}, \dots, m_t^{(M)}\}$. We designate all states of the chain that possess the value i of the component i_t as *level* i of the Markov chain ξ_t.

We shall denote the infinitesimal generator of this chain as Q. The matrix Q encompasses the intensities of all potential transitions of the examined chain throughout an infinitesimally short period.

To formulate the expression for the generator Q, we require the subsequent auxiliary result that delineates the behavior of the M-dimensional vector random process $\mathbf{m}_t = \{m_t^{(1)}, \dots, m_t^{(M)}\},\ t \geq 0$, given a fixed value n of the component n_t of the Markov chain $\xi_t,\ n = \overline{1, N}$.

Lemma 1. *Let the number of customers in service be denoted as n, where n ranges from 1 to N. The matrix L_n delineates the transition intensities of the process $\mathbf{m}_t$ upon the completion of service at one of the servers. The square matrix A_n of size T_n specifies the transition intensities of the process $\mathbf{m}_t$ when service at one of the servers shifts to another phase. The matrix $P_n(\boldsymbol{\beta})$ characterizes the transition probabilities of the process $\mathbf{m}_t$ at the instant a new customer arrives and begins service. The diagonal entries of the diagonal matrix Δ_n specify the exit rates of the process $\mathbf{m}_t$ from the respective state.*

The matrices L_n, A_n, $P_n(\boldsymbol{\beta})$, and Δ_n are calculated using the procedures for matrices of the same designation outlined in [28], pages 106938–106939.

For use in the subsequent sections, the following denotations are used:

- $\text{diag}\{\dots\}$ means the diagonal or block-diagonal matrix of an appropriate dimension with the diagonal entries or blocks listed in the brackets;
- $d = \binom{N+M}{M} = \sum\limits_{n=0}^{N} T_n$;
- $\Gamma^{(\nu)}$ is the diagonal matrix of size d defined as

$$\Gamma^{(\nu)} = \text{diag}\{\underbrace{0, 0, \dots, 0}_{\sum\limits_{n=0}^{R_\nu} T_n}, \underbrace{1, 1, \dots, 1}_{\sum\limits_{n=R_\nu+1}^{N} T_n}\};$$

- O is the zero matrix of the proper dimension. If the dimension is not clear from the context, it may be indicated by the subscript;

- $G^{(\nu)}_{n,n}$ is the matrix defined as

$$G^{(\nu)}_{n,n} = \begin{cases} O_{T_n}, & 0 \le n \le N-K, \\ p \sum\limits_{k=N-n+1}^{K} (D_k)_{\nu,\nu} I_{T_n}, & N-K < n \le N; \end{cases}$$

- $B^{(\nu)} = (B^{(\nu)}_{n,n'})_{n,n'=\overline{0,N}}$ is a block upper-Hessenberg matrix of size $d \times d$ with non-zero blocks:

$$B^{(\nu)}_{0,0} = G^{(\nu)}_{0,0},\ B^{(\nu)}_{n,n} = A_n + \Delta_n + G^{(\nu)}_{n,n},\ n = \overline{1,N},$$

$$B^{(\nu)}_{n,n+k} = (D_k)_{\nu,\nu} \prod_{j=n}^{n+k-1} P_n(\boldsymbol{\beta}),\ n+k \le N,\ k \le K,\ n = \overline{0,N-1},$$

$$B^{(\nu)}_{n,n-1} = L_n,\ n = \overline{1,N};$$

- $\tilde{B}^{(\nu,\nu')} = (\tilde{B}^{(\nu,\nu')}_{n,n'})_{n,n'=\overline{0,N}}$ is a block matrix of size $d \times d$ with the non-zero blocks

$$\tilde{B}^{(\nu,\nu')}_{n,n+k} = (D_k)_{\nu,\nu'} \prod_{j=n}^{n+k-1} P_n(\boldsymbol{\beta}),\ n+k \le N,\ k \le K,\ n = \overline{0,N-1};$$

$$\tilde{B}^{(\nu,\nu')}_{n,n} = \begin{cases} O, & 0 \le n \le N-K, \\ p \sum\limits_{k=N-n+1}^{K} (D_k)_{\nu,\nu'} I_{T_n}, & N-K < n \le N; \end{cases}$$

- $\bar{B}^{(\nu)} = (\bar{B}^{(\nu)}_{n,n'})_{n,n'=\overline{0,N}}$ is a block matrix of size $d \times d$ with the non-zero blocks $\bar{B}^{(\nu)}_{n,n+1} = P_n(\boldsymbol{\beta}),\ n = \overline{0,R_\nu}$;
- $C^{(\nu,\nu')}_{n,k} = \begin{cases} O, & 0 \le n < N-K+k, \\ (1-p)(D_{N-n+k})_{\nu,\nu'} \prod\limits_{j=n}^{N-1} P_n(\boldsymbol{\beta}), & N-K+k \le n < N, \\ (1-p)(D_k)_{\nu,\nu'} I_{T_N}, & n = N. \end{cases}$

The following statement is true.

Theorem 1. *The infinitesimal generator Q of the Markov chain $\xi_t,\ t \ge 0$ has an upper-Hessenberg structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & Q_{0,4} & \cdots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & Q_{1,3} & Q_{1,4} & \cdots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & Q_{2,4} & \cdots \\ O & O & Q_{3,2} & Q_{3,3} & Q_{3,4} & \cdots \\ O & O & O & Q_{4,3} & Q_{4,4} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The non-zero blocks $Q_{i,j},\ i \ge 0,\ j = \overline{\max\{0, i-1\}, i+K}$ are the square matrices of dimension Vd, which are defined as follows:

1. $Q_{i,i} = (Q^{(\nu,\nu')}_{i,i})_{\nu,\nu'=\overline{1,V}}, i \ge 0$, *where*

$$Q^{(\nu,\nu)}_{i,i} = (D_0)_{\nu,\nu} I_d + B^{(\nu)} - i(\gamma + \alpha) I_d + i\alpha\Gamma^{(\nu)},\ \nu = \overline{1,V},$$

$$Q^{(\nu,\nu')}_{i,i} = \tilde{B}^{(\nu,\nu')} + (D_0)_{\nu,\nu'} I_d,\ \ \nu, \nu' = \overline{1,V},\ \nu \ne \nu';$$

2. $Q_{i,i-1} = \text{diag}\{Q_{i,i-1}^{(\nu,\nu)},\ \nu = \overline{1,V}\},\ i \geq 1,$ *where*

$$Q_{i,i-1}^{(\nu,\nu)} = i\gamma I_d + i\alpha \bar{B}^{(\nu)};$$

3. $Q_{i,i+k} = (Q_{i,i+k}^{(\nu,\nu')})_{\nu,\nu'=\overline{1,V}}, i \geq 0, k = \overline{0,K},$ *where the blocks* $Q_{i,i+k}^{(\nu,\nu')}$ *have the form*

$$Q_{i,i+k}^{(\nu,\nu')} = \begin{pmatrix} O & \dots & O & C_{0,k}^{(\nu,\nu')} \\ O & \dots & O & C_{1,k}^{(\nu,\nu')} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & C_{N,k}^{(\nu,\nu')} \end{pmatrix}.$$

Proof. The theorem is established by examining the intensities of all potential transitions of the Markov chain ξ_t during an infinitesimal time interval. Since during such a period, customers may enter the orbit in groups of up to K customers and leave it one at a time, the matrices $Q_{i,j},\ i,j \geq 0$ are zero matrices for all i,j such that $i-1 > j > i+K,\ i \geq 0$.

The matrices $Q_{i,i},\ i \geq 0$ consist of the blocks $Q_{i,i}^{(\nu,\nu')},\ \nu,\nu' = \overline{1,V}$.

The diagonal elements of the diagonal blocks $Q_{i,i}^{(\nu,\nu)}$ are negative and their absolute values determine the exit rates of the Markov chain ξ_t from its states. The Markov chain ξ_t may transition from its current state under the four following circumstances:

(a) The fundamental process ν_t of arrivals transitions from the present state ν, with transition rates determined by the absolute values of the diagonal entries of $(D_0)_{\nu,\nu} I_d$.

Note that the exit of the Markov chain ξ_t from its state does not occur when the fundamental process ν_t makes a transition from the state ν to the state ν with the generation of a batch of customers, but this batch is lost due to a lack of a sufficient number of idle servers. This explains the presence of matrices $G_{n,n}^{(\nu)}$, which define such intensities, in the expression for the summand $B^{(\nu)}$ in the formula for the block $Q_{i,i}^{(\nu,\nu)}$.

(b) A customer departs the orbit due to impatience (the transition rates are represented by the diagonal elements of the matrix $i\gamma I_d$).

(c) A customer departs the orbit due to a successful retry attempt. The transition rates are represented by the diagonal elements of the matrix $i\alpha I_d - i\alpha\Gamma^{(\nu)}$. Note that the exit of the Markov chain ξ_t from its state does not happen if a customer attempts to access service, provided that the number of occupied servers exceeds the threshold R_ν. In this case, the attempt is not successful, and a customer returns to orbit. Consequently, this explains the presence of the supplementary term $i\alpha\Gamma^{(\nu)}$, which defines the intensities of unsuccessful attempts in the diagonal blocks of the generator.

(d) the fundamental process $\mathbf{m}_t,\ t \geq 0$ of service departures from its present state. The transition rates in this case are specified by the matrix $\text{diag}\{0, \Delta_n,\ n = \overline{1,N}\}$. This explains the presence of the diagonal matrices $\Delta_n,\ n = \overline{1,N}$ in summand $B^{(\nu)}$.

The non-diagonal elements of the matrices $Q_{i,i}^{(\nu,\nu)}$ dictate the transition rates of the Markov chain ξ_t while preserving the values of the components i and ν. The transitions are characterized by the non-diagonal elements of the matrix $B^{(\nu)}$. The non-diagonal elements of the matrix $B^{(\nu)}$ represent the intensities of the following transitions:

(a) The fundamental process of service makes a transition without completing the service in one of the occupied servers (transition rates are specified by the elements of the matrix $\text{diag}\{0, A_n,\ n = \overline{1,N}\}$).

(b) The fundamental process ν_t transitions from state ν to state ν while generating a batch of customers, with the batch size being less than or equal to the number of idle servers.

In this case, all customers from the batch start service. The relevant intensities are included in the blocks $B^{(\nu)}_{n,n+k}$, which depend on the number of busy servers and the size of the incoming batch.

(c) The service is finished in one of the busy servers (transition rates are defined by the elements of the matrix $\text{diag}^{-}\{L_n,\ n = \overline{1,N}\}$, where diag^{-} is a block sub-diagonal matrix).

The matrices $Q^{(\nu,\nu')}_{i,i}$, where $\nu = \overline{1,V}$, $\nu \neq \nu'$, define the transition rates of the Markov chain ξ_t that change the state of the arrival process but maintain the number of customers in the orbit. These matrices represent the summation of two matrices $\tilde{B}^{(\nu,\nu')}$ and $(D_0)_{\nu,\nu'} I_d$. The matrix $\tilde{B}^{(\nu,\nu')}$ delineates the rates of the fundamental process ν_t transition from state ν to state ν' with the following:

(i) The generation of a batch, the size k of which does not exceed the number of idle servers. All arriving customers begin service. The transition probabilities of the process $\mathbf{m}_t$ at this time are represented by the matrix $\prod\limits_{j=n}^{N-1} P_n(\boldsymbol{\beta})$.

(ii) The generation of a batch, the size of which exceeds the number of idle servers. This batch abandons the system.

The matrix $(D_0)_{\nu,\nu'} I_d$ defines the fundamental process ν_t transition rates from the state ν to the state ν' without batch generation.

Therefore, we obtain the view of the blocks $Q_{i,i},\ i \geq 0$.

The Markov chain ξ_t can transition from level i to level $i-1$ only if (a) one of the i customers from the orbit is lost due to impatience (the number of busy servers remains unchanged) or (b) one of the i customers in the orbit successfully enters service (the number of busy servers increases by one). The matrices $Q_{i,i-1}$ have only diagonal blocks $Q^{(\nu,\nu)}_{i,i-1}$, $\nu = \overline{1,V}$ because the fundamental process ν_t of the $BMAP$ cannot make a transition during the moment of retrials or a customer departure due to impatience. Consequently, in scenario (a), the intensities are represented by the matrix $i\gamma I_d$, whereas, in scenario (b), the relevant intensities correspond to the elements of the matrix $i\alpha\bar{B}^{(\nu)}$.

Transitions of the Markov chain ξ_t from level i to level $i+k$ occur exclusively due to the transitions of the fundamental process ν_t of the $BMAP$ between states ν and ν' under the following conditions: (a) the number of occupied servers equals N and the incoming batch of k customers remains in the system, joining the orbit, or (b) the number of available servers is insufficient for the incoming batch, resulting in part of the batch occupying the idle servers while the remaining k customers proceed to the orbit. The matrix $(1-p)(D_k)_{\nu,\nu'} I_{T_N}$ delineates the transition intensities for case (a), while the matrix $(1-p)(D_{N-n+k})_{\nu,\nu'} \prod\limits_{j=n}^{N-1} P_n(\boldsymbol{\beta})$ encompasses the transition intensities for case (b). Taking these reasonings into account, we obtain the form of the blocks $Q_{i,j},\ i,j \geq 0,\ j \leq i+K$.

We define the view of all blocks of the generator; thus, the theorem is proven. □

Given that the blocks $Q_{i,i}$ and $Q_{i,i-1}$ of the generator are explicitly contingent on the level i, the Markov chain ξ_t clearly does not fall within the category of the $M/G/1$-type Markov chain; see [29]. This makes the analysis of this chain difficult.

It can be confirmed that the following matrices $Y_k,\ k \geq 0$, exist:

$$Y_k = \lim_{i\to\infty} \mathcal{Z}_i^{-1} Q_{i,i+k-1} + \delta_{k,1} I,\ k \geq 0,$$

where $\delta_{k,1}$ is the so-called Kronecker delta: $\delta_{k,1} = 1$ if $k = 1$ and $\delta_{k,1} = 0$ otherwise, and the matrix $\mathcal{Z}_i$ is defined by the formula $\mathcal{Z}_i = -I \circ Q_{i,i}$, where $\circ$ is the Hadamard product of the matrices symbol; see, e.g., [30]. The matrix $\sum\limits_{k=0}^{\infty} Y_k$ is the stochastic one. This indicates that the Markov chain ξ_t is classified as an Asymptotically Quasi-Toeplitz Markov Chain (AQTMC), as introduced in [31] and further discussed in [16]. In [31], sufficient requirements for the ergodicity of such chains and a numerically stable algorithm for calculating their stationary distribution are presented.

The direct application of the results of [31] requires the derivation of the explicit form of the matrices $Y_k,\ k \geq 0$. However, it is possible to avoid this work because we assume that customers residing in the orbit are impatient ($\gamma > 0$). Applying the result from [32], we conclude that the following assertion is true.

Lemma 2. *Due to the impatience of customers staying in the orbit ($\gamma > 0$), the Markov chain ξ_t is ergodic for every set of system parameters.*

Therefore, there exist the stationary probabilities of these Markov chain states:

$$\pi(i,\nu,n,m^{(1)},\ldots,m^{(M)}) =$$

$$\lim_{i\to\infty} P\{i_t = i, \nu_t = \nu, n_t = n, m_t^{(1)} = m^{(1)},\ldots,m_t^{(M)} = m^{(M)}\},$$

$$i \geq 0,\ \nu = \overline{1,V},\ n = \overline{0,N},\ m^{(l)} = \overline{0,n},\ l = \overline{1,M},\ \sum_{l=1}^{M} m^{(l)} = n.$$

Utilizing the aforementioned enumeration of the states of the Markov chain ξ_t, we indicate by $\boldsymbol{\pi}_i$ the row vectors representing the stationary probability of the states of the Markov chain corresponding to level $i,\ i \geq 0$.

The vectors $\boldsymbol{\pi}_i$ can be partitioned as $\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \ldots, \boldsymbol{\pi}(i,N))$, where $\boldsymbol{\pi}(i,n) = (\boldsymbol{\pi}(i,n,1), \boldsymbol{\pi}(i,n,2), \ldots, \boldsymbol{\pi}(i,n,W)),\ n = \overline{0,N}$.

The vectors $\boldsymbol{\pi}_i,\ i \geq 0$ are established to fulfill the system of linear algebraic equations:

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

where Q is the generator of the $ACTMC$ $\xi_t,\ t \geq 0$.

Due to the upper-Hessenberg structure of the matrix Q, it is possible to obtain from this system an expression for the unknown vectors $\boldsymbol{\pi}_i$ in the form $\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \Phi_i,\ i \geq 1$, where the matrices $\Phi_i,\ i \geq 1$ are computed recursively. However, the recursion for the matrices $\Phi_i,\ i \geq 1$ lacks numerical stability, and it remains ambiguous how to determine the unknown vector $\boldsymbol{\pi}_0$. The problem of calculating this vector is quite difficult even in the case when the upper-Hessenberg structure of the matrix Q is additionally quasi-Toeplitz, i.e., the value of the blocks $Q_{i,j}$ depends only on the difference between i and j, but not on i and j individually. In this case, certain additional probabilistic considerations can be used; see [29]. Alternatively, the equation for the vector $\boldsymbol{\pi}_0$ is derived using the reasonings of the analyticity of the vector-generating function of the vectors $\boldsymbol{\pi}_i,\ i \geq 0$ in the unit disk of the complex plane; see, e.g., [33–35]. In the case of the absence of the quasi-Toeplitz property of the generator, the technique exploiting the analyticity does not work. Therefore, in [31], system (1) for vectors $\boldsymbol{\pi}_i,\ i \geq 0$, is replaced with another system that is derived via the construction of the infinite sequence of so-called censoring Markov chains; see, e.g., [36] with different censoring levels. The corresponding algorithm for solving this alternative system, the solution of which coincides with the solution of system (1), is presented in [31]. Another algorithm that does not need analytical derivations to obtain the explicit form of the matrices $Y_k,\ k \geq 0$ is described in [37]. This algorithm shows a higher convergence rate compared to [31] in the technique for calculating the matrices that delineate transition probabilities of the finite components of the chain ξ_t during the initial passage time to the level i starting from the level $i+1$.

4. System Performance Characteristics

Having calculated the vectors of stationary probabilities of the system states, we can determine the values of various probabilistic characteristics of the system.

The mean quantity L_{system} of customers in the system (including staying in the orbit and receiving service in servers) is

$$L_{system} = \sum_{i=0}^{\infty} \sum_{\nu=1}^{V} \sum_{n=0}^{N} (i+n)\boldsymbol{\pi}(i,\nu,n)\mathbf{e}_{T_n}.$$

The average quantity L_{orbit} of customers present at a given moment in the orbit is

$$L_{orbit} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i \mathbf{e}_{dV}.$$

The mean quantity N_{server} of busy servers is computed as

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

The presented formulas for L_{system}, L_{orbit}, and N_{server} are evidently derived as the expectations of the corresponding discrete random variables.

The probability that all servers are idle at an arbitrary moment is

$$P_{idle-servers} = \sum_{i=0}^{\infty} \sum_{\nu=1}^{V} \boldsymbol{\pi}(i,\nu,0).$$

The probability that the orbit is empty at an arbitrary moment is

$$P_{empty-orbit} = \boldsymbol{\pi}_0 \mathbf{e}_{dV}.$$

The probability that the system is empty at an arbitrary moment is

$$P_{empty-system} = \sum_{\nu=1}^{V} \boldsymbol{\pi}(0,\nu,0).$$

The average intensity λ_{out} of the output flow of serviced customers is

$$\lambda_{out} = \sum_{i=0}^{\infty} \sum_{\nu=1}^{V} \sum_{n=1}^{N} \boldsymbol{\pi}(i,\nu,n)L_n \mathbf{e}_{T_{n-1}}.$$

This formula evidently follows from the formula of total probability, taking into account that the output rate of serviced customers is equal to $L_n\mathbf{e}_{T_{n-1}}$ when the number of customers in the orbit is equal to $i,\ i \geq 0$, the state of the fundamental process of arrivals is equal to $\nu,\ \nu = \overline{1,V}$, and the number of busy servers is equal to $n,\ n = \overline{1,N}$.

The probability $P_{to-orbit}$ that an arbitrary arriving customer joins the orbit is

$$P_{to-orbit} = \frac{1-p}{\lambda} \sum_{i=0}^{\infty} \sum_{\nu=1}^{V} \sum_{\nu'=1}^{V} \sum_{n=0}^{N} \sum_{k=N-n+1}^{K} (k-(N-n))(D_k)_{\nu,\nu'}\boldsymbol{\pi}(i,\nu,n)\mathbf{e}_{T_n}.$$

This probability is calculated as the ratio of the rate of arrival of customers who join the orbit to the average arrival rate λ. The former rate is calculated as follows. Let the number of customers in the orbit be i and the number of busy servers be n. The instantaneous rate of arrival of customers in batches of size k is equal to $k(D_k)_{\nu,\nu'}$ when the fundamental process of arrivals makes a jump from the state ν to the state ν'. An arriving customer has a chance to join the orbit if the size k of the batch exceeds the number $N-n$ of idle servers. For an arbitrary customer arriving in a batch of size k, it is natural to assume that the customers are enumerated in a random manner, the customer receives the lth position in the batch with the probability $\frac{1}{k}$ for all l from 1 to k, and only the suitable number of customers from the head of a batch can be admitted to the orbit (with the probability $(1-p)$). As the result of these considerations, it is easy to see that the rate of arrival of customers who join the orbit is calculated as $(1-p)\sum\limits_{i=0}^{\infty}\sum\limits_{\nu=1}^{V}\sum\limits_{\nu'=1}^{V}\sum\limits_{n=0}^{N}\sum\limits_{k=N-n+1}^{K} k(D_k)_{\nu,\nu'}\frac{(k-(N-n))}{k}\boldsymbol{\pi}(i,\nu,n)\mathbf{e}_{T_n}$. Thus, we obtain the expression under the proof for computation of the probability $P_{to-orbit}$.

The proof of several subsequent formulas is similar.

The probability $P_{arrloss}$ that an arbitrary customer is lost upon arrival due to all servers being busy is computed as

$$P_{arrloss} = \frac{p}{\lambda}\sum_{i=0}^{\infty}\sum_{\nu=1}^{V}\sum_{\nu'=1}^{V}\sum_{n=0}^{N}\sum_{k=N-n+1}^{K} k(D_k)_{\nu,\nu'}\boldsymbol{\pi}(i,\nu,n)\mathbf{e}_{T_n}.$$

The probability $P_{arrloss}^{(\nu)}$ that an arbitrary customer is lost upon arrival due to an insufficient number of idle servers, conditional on the fundamental process ν_t of the $BMAP$ being in the state ν, is computed as

$$P_{arrloss}^{(\nu)} = \frac{p}{\lambda}\sum_{i=0}^{\infty}\sum_{\nu'=1}^{V}\sum_{n=0}^{N}\sum_{k=N-n+1}^{K} k(D_k)_{\nu,\nu'}\boldsymbol{\pi}(i,\nu,n)\mathbf{e}_{T_n}.$$

The probability $P_{imm-access}$ that an arbitrary customer immediately starts service upon arrival is

$$P_{imm-access} = \frac{1}{\lambda}\sum_{i=0}^{\infty}\sum_{\nu=1}^{V}\sum_{\nu'=1}^{V}\sum_{n=0}^{N-1}\Bigg[\sum_{k=1}^{\min\{k,N-n\}} k(D_k)_{\nu,\nu'}\boldsymbol{\pi}(i,\nu,n)\mathbf{e}_{T_n}+$$

$$(1-p)\sum_{k=N-n+1}^{K}(N-n)(D_k)_{\nu,\nu'}\boldsymbol{\pi}(i,\nu,n)\mathbf{e}_{T_n}\Bigg].$$

The probability $P_{imploss}$ that an arbitrary customer is lost from the orbit due to impatience is

$$P_{imploss} = \frac{\gamma}{\lambda}\sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}_{dV} = \frac{\gamma L_{orbit}}{\lambda}.$$

This formula is clear because the probability $P_{imploss}$ of an arbitrary customer loss from the orbit due to impatience is the ratio of the rate of the lost due to impatience customers γL_{orbit} to the customer arrival rate.

The probability $P_{imploss}^{(\nu)}$ that an arbitrary customer is lost from the orbit due to impatience when the fundamental process ν_t of the $BMAP$ is in the state ν is computed as

$$P_{imploss}^{(\nu)} = \frac{\gamma}{\lambda}\sum_{i=1}^{\infty}\sum_{n=0}^{N} i\boldsymbol{\pi}(i,\nu,n)\mathbf{e}_{T_n}.$$

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

$$P_{loss} = 1 - \frac{\lambda_{out}}{\lambda} = P_{arrloss} + P_{imploss}.$$

This formula, which delineates two distinct methods for calculating P_{loss}, is crucial in ensuring the accuracy of the computations for the generator Q and the stationary distribution $\boldsymbol{\pi}_i,\ i \geq 0$ of the system under examination.

5. Numerical Examples

In this section, we present the results of two numerical experiments.

Experiment 1. The purpose of Experiment 1 is to numerically highlight the dependence of the main performance measures of the system on the parameters R_ν, $\nu = \overline{1, W}$ of the control strategy and show the possibility of obtaining the optimal values of these thresholds. To this end, let us assume that customers enter the system in the $BMAP$ defined by the matrices

$$D_0 = \begin{pmatrix} -9.27696 & 0.553551 \\ 0.666133 & -3.65263 \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 4.3081 & 0.34438 \\ 0.0442795 & 1.54852 \end{pmatrix}, D_2 = \begin{pmatrix} 2.15405 & 0.17219 \\ 0.0221398 & 0.77426 \end{pmatrix},$$

$$D_3 = \begin{pmatrix} 1.07703 & 0.0860951 \\ 0.0110699 & 0.38713 \end{pmatrix}, D_4 = \begin{pmatrix} 0.538513 & 0.0430475 \\ 0.00553494 & 0.193565 \end{pmatrix}.$$

In this arrival flow, the customers arrive in batches of the maximal size $K = 4$, with the average arrival rate of groups being $\lambda_g = 5.19231$. The mean customer arrival rate in the $BMAP$ is $\lambda = 9$, the correlation coefficient of two consecutive periods between batch arrivals is 0.0992515, and the coefficient of variation is 1.39728. The vector of the stationary distribution of the fundamental process of the $BMAP$ is $\boldsymbol{\theta} = (0.384494, 0.615506)$. The mean customer arrival rate at state k ($k = 1, 2$) of the fundamental process of the $BMAP$ is the k-th element of the vector $\boldsymbol{\theta} \sum\limits_{k=1}^{K} D_k = (5.47201, 3.52799)$. The conditional mean arrival rates given states 1 and 2 of the fundamental process of the $BMAP$ are equal to 14.23171 and 5.73185, correspondingly.

We assume that the service time distribution in each server is of phase type and is defined by the vector $\boldsymbol{\beta} = (0.5, 0.5)$ and the matrix $S = \begin{pmatrix} -2 & 0 \\ 0 & -\frac{2}{3} \end{pmatrix}$. The average service time of a customer is $b_1 = 1$. The squared coefficient of variation of service time is 4.

We assume that the number of servers is $N = 15$, the retrial intensity is $\alpha = 0.1$, the intensity of impatience is $\gamma = 0.008$, and the probability of loss of a batch in the case of lack of enough idle servers to accommodate all arriving customers is $p = 0.4$.

To construct the 3-D representations depicting the dynamics of several performance metrics of the system, we will adjust the thresholds R_1 and R_2 throughout the range $[0, N-1]$ in increments of 1.

Figures 2 and 3 illustrate the dependencies of the mean quantity L_{orbit} of customers in the orbit, the mean quantity N_{server} of busy servers, and loss probabilities $P^{(1)}_{imploss}$ and $P^{(2)}_{imploss}$ due to impatience on the thresholds R_1 and R_2.

Figure 4 shows the dependencies of the probabilities $P^{(1)}_{arrloss}$ and $P^{(2)}_{arrloss}$ that an arbitrary customer is lost upon arrival due to the lack of available servers on the thresholds R_1 and R_2.

Figures 5 and 6 illustrate the relationship between the loss probabilities $P_{arrloss}$, $P_{imploss}$, and P_{loss} and the varying values of R_1 and R_2.

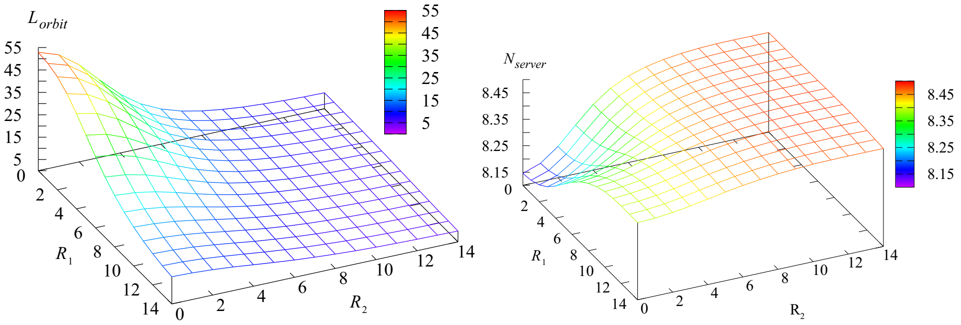

Figure 2. Dependence of L_{orbit} and N_{server} on the parameters R_1 and R_2.

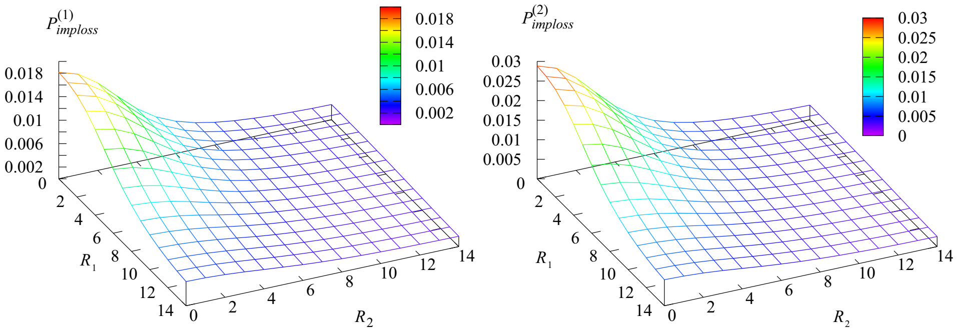

Figure 3. Dependence of $P^{(1)}_{imploss}$ and $P^{(2)}_{imploss}$ on the parameters R_1 and R_2.

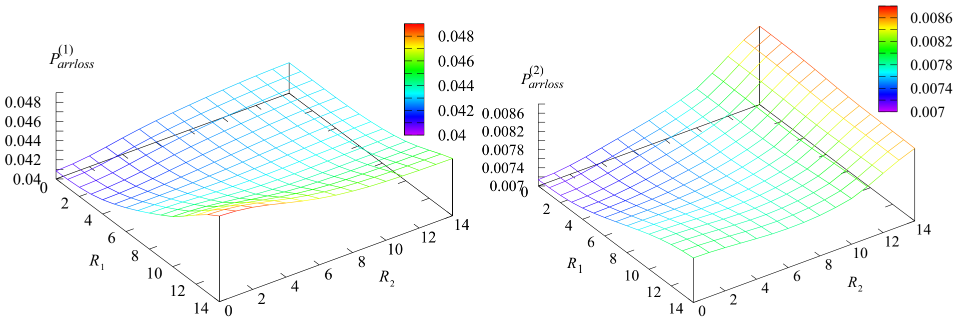

Figure 4. Dependence of $P^{(1)}_{arrloss}$ and $P^{(2)}_{arrloss}$ on the parameters R_1 and R_2.

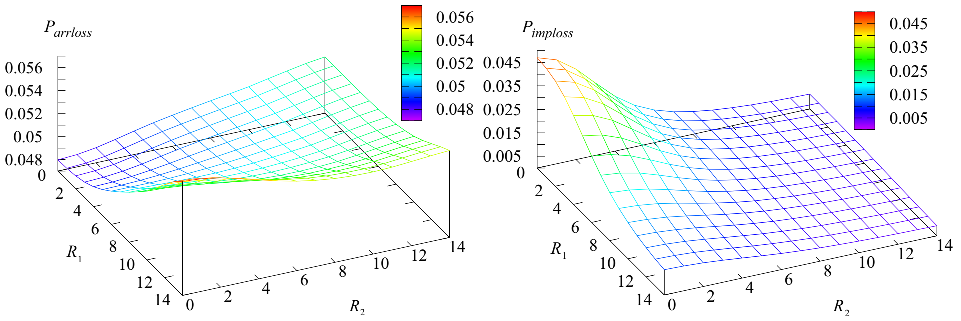

Figure 5. Dependence of $P_{arrloss}$ and $P_{imploss}$ on the parameters R_1 and R_2

As seen from Figure 2, the mean quantity of customers in orbit L_{orbit} takes its maximum value of 52.875 when $R_1 = R_2 = 0$. In this case, the customers from orbit have a chance to start service only if, during the retrial epoch, all servers are free. Also, L_{orbit} decreases with the increase in the parameters R_1 and R_2. The minimal value of L_{orbit} is 4.838 when $R_1 = R_2 = 14$, i.e., the admission control for retrial customers is not applied.

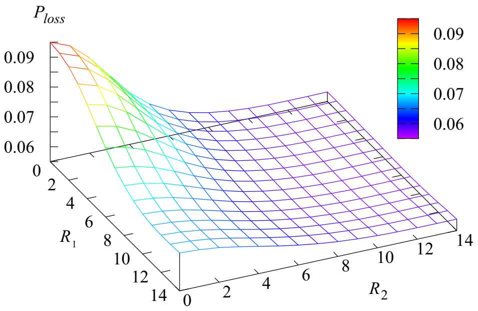

Figure 6. Dependence of P_{loss} on the parameters R_1 and R_2.

The mean quantity of busy servers N_{server} sharply increases with growth in the parameters R_1 and R_2 when R_1 and R_2 are small, e.g., less than 6. This can be explained by the fact that with such growth, the mean quantity of customers in orbit L_{orbit} decreases, which causes the decrease in the loss probabilities of a customer due to impatience $P_{imploss}$, $P^{(1)}_{imploss}$, and $P^{(2)}_{imploss}$; see Figures 3 and 5. The fewer customers are lost, the more customers receive service and the more servers are busy. With further growth in R_1 and R_2, the value of N_{server} grows less essential and then starts to slowly decrease. This is caused by the fact that the growth in R_1 and R_2 causes the increase in the loss probabilities of customers upon arrival $P_{arrloss}$, $P^{(1)}_{arrloss}$, and $P^{(2)}_{arrloss}$; see Figures 4 and 5.

In contrast to the probabilities $P_{imploss}$, $P^{(1)}_{imploss}$, and $P^{(2)}_{imploss}$ of customer loss due to impatience, which decrease with a decrease in R_1 and R_2, the loss probabilities $P_{arrloss}$, $P^{(1)}_{arrloss}$, and $P^{(2)}_{arrloss}$ behave not so predictably. For example, one can anticipate that the loss probability $P_{arrloss}$ takes its maximal value in the case $R_1 = R_2 = 14$ when retrial customers can enter service without any restriction, but this is not true.

In this example, the maximal value of $P_{arrloss}$ is 0.0569 for $R_1 = 14$ and $R_2 = 0$, while $P_{arrloss} = 0.05455$ if $R_1 = R_2 = 14$. This can be explained as follows. Since the average arrival rate is essentially less under state 2 than under state 1, the load of the system is also less under state 2. If we fix $R_2 = 0$, then when the arrival flow is in state 2, the customers from the orbit can start service only if all servers are idle. Thus, a lot of customers accumulate in the orbit, and when the fundamental process of $BMAP$ transits to state 1, a lot of customers from orbit start to compete with primary customers. If no admission control is applied ($R_1 = 14$), a lot of primary customers are lost. This also explains the non-monotonic behavior of some other loss probabilities of a customer upon arrival.

For example, if we fix $R_1 = 14$, $P^{(2)}_{arrloss} = 0.007979$ for $R_2 = 0$, $P^{(2)}_{arrloss} = 0.00786$ for $R_2 = 6$, and $P^{(2)}_{arrloss} = 0.00859$ for $R_2 = 14$. When the admission control under state 2 of the $BMAP$ is not applied ($R_2 = 14$), many arriving primary customers are lost due to a lack of available servers. But, if the admission control is the strictest ($R_2 = 0$), during state 2, retrial customers have a very low chance of entering service (while the arrival rate of primary customers is relatively low and, therefore, a scarcity of servers is possible), and a lot of customers wait in the orbit. After the return of the fundamental process of the $BMAP$ to state 1, customers from the orbit occupy all servers quickly, and it is very likely that a lot of quickly arriving primary customers are rejected.

The total loss probability for a customer P_{loss} is the sum of the probabilities $P_{imploss}$ and $P_{arrloss}$. In this experiment, the probability $P_{imploss}$ is more sensitive to the parameters R_1 and R_2, resulting in P_{loss} exhibiting behavior akin to that of $P_{imploss}$. However, in contrast to $P_{imploss}$, the probability P_{loss} behaves non-monotonically. It takes its minimal value of 0.05752 for $R_1 = 9$ and $R_2 = 12$. The fact that the optimal value of R_1 is less than the optimal value of R_2, i.e., the restriction of retrying customer access is more strict under state 1 of the fundamental process, is intuitively clear. Since the arrival rate of primary customers during state 1 is essentially higher, it makes sense to restrict the access of retrying customers more strongly.

Thus, if the purpose of the system optimization is to minimize the loss probability of customers, the values $R_1 = 9$ and $R_2 = 12$ should be considered as the optimal ones in this example. However, in real-world systems, cost criteria may be more complicated.

We will suppose that the efficacy of the system's operation is assessed using the subsequent cost criterion:

$$J(R_1, R_2) = a\lambda_{out} - b_1\lambda P_{arr-loss} - b_2\lambda P_{imploss},$$

where a is the profit gained by the system for one customer service, b_1 represents the cost associated with the loss of one customer at the system's entrance, and b_2 is the charge for one customer loss due to impatience.

The cost criterion $J(R_1, R_2)$ delineates the average profit of the system per unit of time. The goal is to find the values of R_1 and R_2 that provide the maximum value to this criterion.

The following cost coefficients are set up: $a = 1$, $b_1 = 6$, $b_2 = 3$.

Figure 7 illustrates the relationship between the cost criterion J and the parameters R_1 and R_2.

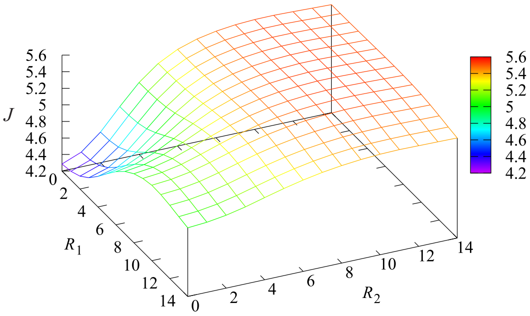

Figure 7. Dependence of the cost criterion J on the parameters R_1 and R_2.

The optimal value of the cost criterion is $J^* = 5.53889$, attained at $R_1^* = 7$ and $R_2^* = 10$.

Experiment 2. This experiment aims to demonstrate the significance of considering the potential of customer arrivals in groups. Consequently, rather than the aforementioned $BMAP$, we examine the MAP arrival flow with equivalent average intensity, coefficients of variation, and correlation. This MAP is characterized by the subsequent matrices:

$$D_0 = \begin{pmatrix} -16.0801 & 0.959488 \\ 1.15463 & -6.33123 \end{pmatrix}, \, D_1 = \begin{pmatrix} 14.0013 & 1.11924 \\ 0.143908 & 5.03269 \end{pmatrix}.$$

The average customer arrival rate is equal to 9. The vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta} D_1$ are the same as the corresponding vectors characterizing the $BMAP$ considered in the first experiment.

Other parameters of the queueing system are assumed to be the same as in the previous experiment.

The dependencies of the loss probabilities $P_{arrloss}$, $P_{imploss}$, and P_{loss} and the cost criterion J on the different values of R_1 and R_2 are shown in Figures 8 and 9.

One can see from Figures 8 and 9 that the consideration of MAP instead of $BMAP$ with an identical average arrival rate leads to overoptimistic results. If the batch arrivals are not taken into account, the loss probability of an arriving customer $P_{arrloss}$ will be more than two times less than its real value. The minimal value of $P_{arrloss}$ is 0.04807 in the case of $BMAP$, while, in the case of MAP, this probability does not exceed 0.02 for any values of R_1 and R_2. The probabilities $P_{imploss}$ and P_{loss} are also essentially less in the case of MAP.

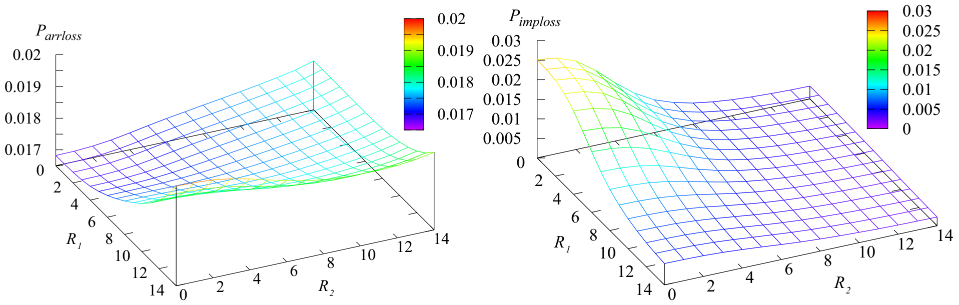

Figure 8. Dependence of $P_{arrloss}$ and $P_{imploss}$ on the parameters R_1 and R_2 for the MAP arrival process.

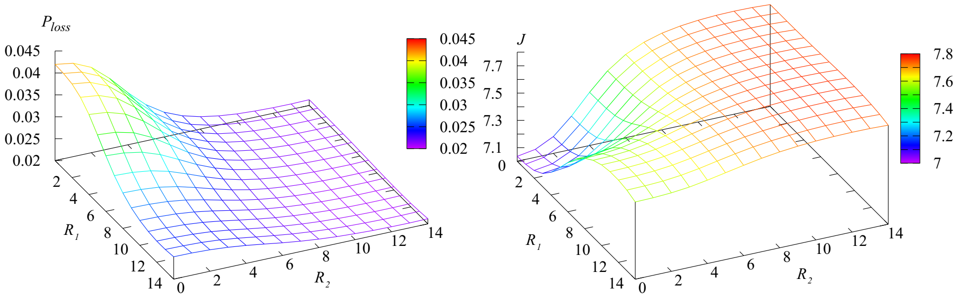

Figure 9. Dependence of P_{loss} and the cost criterion J on the parameters R_1 and R_2 for the MAP arrival process.

The optimal value of the cost criterion in the case of MAP arrival flow is $J^* = 7.7709$ and is achieved for $R_1^* = 9$ and $R_2^* = 12$, about 40 percent higher than in the case of $BMAP$.

It is worth noting that in Experiment 1, we considered the $BMAP$ with an average batch size of 1.7333. Thus, most of the batches were of size 1. It is evident that if someone considers a $BMAP$ with a higher average batch size, the difference between the characteristics of queues with $BMAP$ and MAP arrival models will be even more essential. The conclusions from Experiment 2 prove the importance of the extension of the research, the results of which are presented in [25].

6. Conclusions

In this paper, the $BMAP/PH/N$-type retrial queue with a hybrid of complete rejection/partial admission discipline for primary customer acceptance and server reservation for these customers (or retrials restriction) was analyzed. Retrial customers had access to service solely when the number of occupied servers at the retrial moment was less than or equal to a predetermined threshold, which was dependent on the present state of the primary customers' arrival process. The proper choice of the thresholds allowed for more strict restrictions on the access of retrial customers to servers when the primary customers arrived at a high rate and gave more free access when few primary customers arrived. This allowed for a reduction in the probability of the primary customer's loss upon arrival with a relatively low probability of retrial customer loss due to impatience. To evaluate the performance metrics of the system for predetermined thresholds, it was necessary to calculate the stationary distribution of the Markov chain that characterized the system's behavior. The issue of this chain design was briefly examined, and a description of service in busy servers was provided by considering the number of customers currently receiving service at each level of the fundamental service process. The generated Markov chain was ergodic for all values of the system parameters and thresholds of the admission control due to the impatience of the orbiting customers. It was recommended to use stable algorithms for the computation of the stationary distribution of the Markov chain, which is classified as an asymptotically quasi-Toeplitz Markov chain. The system's key performance characteristics were computed using specific formulas. The results of numerical experiments were presented, which demonstrate the potential for improving the system's operation by selecting the optimal thresholds. The necessity of carefully taking into account the batch arrival was demonstrated.

The presented analysis can be extended to systems with more than one type of primary customer and server reservation for more priority types, retrial customer non-persistence (along or alternatively with their impatience), unreliable work of servers, clearance of servers, orbit, etc.

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

Funding: This research received support by the Belarusian Republican Foundation for Fundamental Research of the Republic of Belarus (grant F25UZB-016) and the Ministry of Higher Education, Science and Innovations of the Republic of Uzbekistan (grant FL-8824063218).

Data Availability Statement: The original contributions presented in this study are included in the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Artalejo, J.R.; Gomez-Corral, A. *Retrial Queueing Systems: A Computational Approach;* Springer-Verla: Berlin/Heidelberg, Germany, 2008.
2. Cohen, J.W. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommun. Rev.* **1957**, *18*, 49.
3. Falin, G. A survey of retrial queues. *Queueing Syst.* **1990**, *7*, 127–167. [CrossRef]
4. Templeton, J.G.; Falin, G.I. *Retrial Queues*; Routledge: London, UK, 2023.
5. Gomez-Corral, A. A bibliographical guide to the analysis of retrial queues through matrix analytic techniques. *Ann. Oper. Res.* **2006**, *141*, 163–191. [CrossRef]
6. Kulkarni, V.G.; Liang, H.M. Retrial queues revisited. *Front. Queueing Model. Appl. Sci. Eng.* **1997**, *7*, 18–34.
7. Yang, T.; Templeton, J.G.C. A survey on retrial queues. *Queueing Syst.* **1987**, *2*, 201–233. [CrossRef]
8. Kim, J.; Kim, B. A survey of retrial queueing systems. *Ann. Oper. Res.* **2016**, *247*, 3–36. [CrossRef]
9. Stepanov, S.; Stepanov, M. Estimation of the performance measures of a group of servers taking into account blocking and call repetition before and after server occupation. *Mathematics* **2021**, *9*, 2811. [CrossRef]
10. Melikov, A.; Chakravarthy, S.R.; Aliyeva, S. A retrial queueing model with feedback. *Queueing Model Serv. Manag.* **2023**, *6*, 63–95.
11. Neuts, M.F. A versatile Markovian point process. *J. Appl. Prob.* **1979**, *16*, 764–779. [CrossRef]
12. Lucantoni, D. New results on the single server queue with a batch Markovian arrival process. *Commun. Stat.-Stoch. Model.* **1991**, *7*, 1–46. [CrossRef]
13. Chakravarthy, S.R. The batch Markovian arrival process: A review and future work. *Adv. Probab. Theory Stoch. Process.* **2001**, *1*, 21–49.
14. Chakravarthy, S.R. *Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach—Basics*; ISTE Ltd.: London, UK; John Wiley and Sons: New York, NY, USA, 2022.
15. Chakravarthy, S.R. *Introduction to Matrix-Analytic Methods in Queues 2: Analytical and Simulation Approach—Queues and Simulation*; ISTE Ltd.: London, UK; John Wiley and Sons: New York, NY, USA, 2022.
16. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer Nature: Berlin/Heidelberg, Germany, 2020; ISBN 978-3-030-32072-0.
17. Gonzalez, M.; Lillo, R.E.; Ramirez-Cobo, J. Call center data modeling: A queueing science approach based on Markovian arrival processes. *Qual. Technol. Quant. Manag.* **2024**. [CrossRef]
18. Buchholz, P.; Kriege, J.; Felko, I. *Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2014.
19. Casale, G.; Zhang, E.Z.; Smirni, E. Trace data characterization and fitting for Markov modeling. *Perform. Eval.* **2010**, *67*, 61–79. [CrossRef]

20. He, Q.M.; Li, H.; Zhao, Y.Q. Ergodicity of the $BMAP/PH/s/s+K$ retrial queue with PH-retrial times. *Queueing Syst.* **2000**, *35*, 323–347. [CrossRef]
21. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*; Courier Corporation: North Chelmsford, MA, USA, 1994.
22. O'Cinneide, C.A. Characterization of phase-type distributions. *Stoch. Model.* **1990**, *6*, 1–57. [CrossRef]
23. Asmussen, S. *Applied Probability and Queues*; Springer: New York, NY, USA, 2003.
24. Breuer, L.; Dudin, A.; Klimenok, V. A retrial $BMAP/PH/N$ system. *Queueing Syst.* **2002**, *40*, 433–457. [CrossRef]
25. Dudin, A.; Krishnamoorthy, A.; Dudin, S.; Dudina, O. Queueing system with control by admission of retrial requests depending on the number of busy servers and state of the underlying process of Markov arrival process of primary requests. *Ann. Oper. Res.* **2024**, *335*, 135–150. [CrossRef]
26. He, Q.M.; Alfa, A.S. Space reduction for a class of multidimensional Markov chains: A summary and some applications. *INFORMS J. Comput.* **2018**, *30*, 1–10. [CrossRef]
27. Ramaswami, V.; Lucantoni, D.M. Algorithms for the multi-server queue with phase type service. *Stoch. Model.* **1985**, *1*, 393–417. [CrossRef]
28. Kim, C.; Dudin, A.; Dudin, S.; Dudina, O. Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users. *IEEE Access* **2021**, *9*, 106933–106946. [CrossRef]
29. Neuts, M.F. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*; CRC Press: Boca Raton, FL, USA, 2021.
30. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 2012.
31. 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]
32. Dudin, A.N.; Dudin, S.A.; Klimenok, V.I.; Dudina, O.S. Stability of queueing systems with impatience, balking and non-persistence of customers. *Mathematics* **2024**, *12*, 2214. [CrossRef]
33. Gail, H.R.; Hantler, S.L.; Taylor, B.A. Non-skip-free M/G/1 and G/M/1 type Markov chains. *Adv. Appl. Probab.* **1997**, *29*, 733–758. [CrossRef]
34. Gail, H.R.; Hantler, S.L.; Taylor, B.A. Spectral analysis of M/G/1 and G/M/1 type Markov chains. *Adv. Appl. Probab.* **1996**, *28*, 114–165. [CrossRef]
35. Gail, H.R.; Hantler, S.L.; Sidi, M.; Taylor, B.A. Linear independence of root equations for M/G/1 type Markov chains. *Queueing Syst.* **1995**, *20*, 321–339. [CrossRef]
36. Kemeny, J.G.; Snell, J.L.; Knapp, A.W. *Denumerable Markov Chains: With a Chapter of Markov Random Fields by David Griffeath*; Springer Science Business Media: Berlin/Heidelberg, Germany, 2012.
37. 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]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.