mathematics

MDPI

Article

Stability of Queueing Systems with Impatience, Balking and Non-Persistence of Customers

Alexander N. Dudin *, Sergey A. Dudin, Valentina I. Klimenok and Olga S. Dudina

Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus; vklimenok@yandex.ru (V.I.K.)
* Correspondence: dudin@bsu.by

Abstract: The operation of many queueing systems is adequately described by the structured multidimensional continuous-time Markov chains. The most well-studied classes of such chains are level-independent Quasi-Birth-and-Death processes, $GI/M/1$ type and $M/G/1$ type Markov chains, generators of which have the block tri-diagonal, lower- and upper-Hessenberg structure, respectively. All these classes assume that the matrices of transition rates are quasi-Toeplitz. This property greatly simplifies their analysis but makes them inappropriate for the study of many important systems, e.g., retrial queues with a retrial rate depending on the number of customers in orbit, queues with impatient customers, etc. The importance of such systems attracts significant interest to their analysis. However, in the literature, there is a methodological gap relating to the ergodicity condition of the corresponding Markov chains. To fulfill this gap and facilitate the analysis of a wide range of such systems, we show that under non-restrictive assumptions, the following hold true: (i) if the customers can balk or are impatient or non-persistent, then the Markov chain describing the behavior of the system belongs to the class of asymptotically quasi-Toeplitz Markov chains; (ii) this chain is ergodic; (iii) known algorithms can be applied for the calculation of the stationary distribution of the corresponding queueing system.

Keywords: ergodicity; multidimensional continuous-time asymptotically quasi-Toeplitz Markov chain; impatience; retrials

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

Citation: 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. https://doi.org/10.3390/math12142214

Academic Editor: Manuel Alberto M. Ferreira

Received: 27 June 2024
Revised: 11 July 2024
Accepted: 14 July 2024
Published: 15 July 2024

1. Introduction

The impatience (reneging, abandonment, etc.) of customers is a typical feature of many (if not all) real-world service systems. If the processing of a customer that waits in the queue does not start during a certain period of time (patience time), the customer leaves the system, independently of other customers, without service. Impatience is related to psychological reasons if the customers are humans, the obsolescence of information, perishing of the products, departure of the waiting mobile user from the cell, etc. As early works on queues with customers' impatience, the papers [1–5] can be mentioned. Now, the literature devoted to the analysis of queues with impatient customers is very extensive. Results of research in the field of queues with customer impatience are presented, e.g., in [6–9].

If the total rate of customer departure due to impatience increases with growth in the number of customers in the buffer, it is intuitively clear that the number of customers in the buffer never becomes infinite. This means that such a system is always stable. However, to the best of our knowledge, these intuitive reasonings are not supported by formal statements proven in the existing literature. Here, we present such a proof for a rather wide class of queueing models.

The motivation for writing this paper is twofold. On one hand, in recent times, we have reviewed a lot of papers where the authors consider queues with customer impatience

and completely ignore a discussion of ergodicity conditions or prove it in a non-correct way. On the other hand, we have done our own intensive research on queues with customer impatience, and usually, the reviewers of our manuscripts require the full formal proof of ergodicity of the considered Markov chain. Therefore, we believe that the results presented in this paper will be helpful to researchers. These results remove the problem of the proof of ergodicity and also present the recommendation of numerically stable algorithms for the computation of the stationary distribution of the system states.

The structure of the text is the following. Section 2 contains more detailed motivation for this paper's preparation and some necessary preliminary information from the theory of structured multidimensional Markov chains (MCs). In Section 3, the ergodicity of the queueing system with impatient customers is proved for the system, the counterpart of which with patient customers is described by the Quasi-Birth-and-Death (QBD) process or $M/G/1$ type MC. The ergodicity in the case of customers balking with the probability of joining the queue approaching zero when the queue length increases is also proven there. In Section 4, the ergodicity of the multi-server retrial queueing system with impatient customers is stated. In Section 5, the ergodicity of the multi-server retrial queueing system with non-persistent customers is stated. Section 6 concludes the paper.

2. Preliminary Information

2.1. Basic Information about Multidimensional Continuous-Time Markov Chains under Study

The operation of many queueing systems can be described by a suitably constructed multidimensional continuous-time MC $\{i_t, n_t\}$, $t \geq 0$. The first component of the MC i_t has a countable state space, $i_t \geq 0$, and corresponds to the current number of customers in the queueing system, buffer, orbit or network. The process n_t describes the transitions of a finite component of the MC. The process n_t having a finite state space indeed may be a whole finite set of finite components representing various auxiliary processes, e.g., the number of busy or broken servers in multi-server systems, the state of the underlying process of arrivals (if the arrivals occur in the Markov arrival process (MAP) or batch Markov arrival process ($BMAP$) or marked Markov arrival process ($MMAP$), etc.), the state of the underlying process of service (if the service time has a phase type, PH distribution, or service is defined by the Markov service process, MSP), the random environment, which has an impact on the system operation or the number of customers at other stages of a tandem system with finite intermediate buffers, etc. An account of the physical meaning of these components is very important for writing down the generator of the MC. However, for the purposes of this paper, we assume that the values of these finite components are enumerated in some order. Thus, without the loss of generality, for the simplicity of denotations, in Section 2, we consider the case of only one finite component n_t. In Sections 3 and 4 devoted to retrial queues, we consider the case where the finite component n_t is two-dimensional.

The set of states of the MC having the fixed value, say, i of the first, countable, component is called level i of the MC, $i \geq 0$. We suppose that there exists a finite number i^0, $i^0 \geq 0$, such that the cardinalities of all levels i such that $i > i^0$ are equal. In particular, in Section 2, we will assume that there exists an integer number N such that the component n_t of the MC admits, for any t, $t \geq 0$, and all $i_t \geq i^0$, the values in the set $\{0, 1, 2, \ldots, N\}$. The cardinality of the levels i for $i \geq i^0$ is equal to $N + 1$. The levels having numbers $0, 1, \ldots, i_0 - 1$ can have various dimensions, and we do not impose any specific assumptions about the behavior of the MC for these levels except the obvious requirement of the boundedness of the transition rates from these levels.

Let $\mathcal{Q}$ be the generator of the MC $\{i_t, n_t\}$, $t \geq 0$. Within this paper, we assume that this generator has the upper-Hessenberg structure

$$\mathcal{Q} = \begin{pmatrix} \mathcal{Q}_{0,0} & \mathcal{Q}_{0,1} & \mathcal{Q}_{0,2} & \mathcal{Q}_{0,3} & \mathcal{Q}_{0,4} & \cdots \\ \mathcal{Q}_{1,0} & \mathcal{Q}_{1,1} & \mathcal{Q}_{1,2} & \mathcal{Q}_{1,3} & \mathcal{Q}_{1,4} & \cdots \\ O & \mathcal{Q}_{2,1} & \mathcal{Q}_{2,2} & \mathcal{Q}_{2,3} & \mathcal{Q}_{2,4} & \cdots \\ O & O & \mathcal{Q}_{3,2} & \mathcal{Q}_{3,3} & \mathcal{Q}_{3,4} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (1)$$

where the matrices $\mathcal{Q}_{i,j}$ consist of the entries $\mathcal{Q}_{i,j}^{(n,n')}$ defining, except the diagonal entries $\mathcal{Q}_{i,i}^{(n,n)}$, transition rates from the state (i,n) to the state (j,n'). The mentioned diagonal entries are negative. Their moduli define the rates of the *MC* η exit from the corresponding state. For $i > i^0,\ j \geq i-1,\ \mathcal{Q}_{i,j}$ are square matrices of size $N+1$. Here, O denotes zero matrix. By I, we will denote the identity matrix. If it is necessary, the size of the matrix can be indicated by the subscript.

The popular partial case suggests that the matrix $\mathcal{Q}$ is block tri-diagonal. The *MC*s having a generator of a block tri-diagonal structure are called Quasi-Birth-and-Death processes (*QBD*).

2.2. Level-Independent Quasi-Birth-and-Death Processes and M/G/1 Type Markov Chains

The most well studied in the literature classes of *MC*s having the structure of form (1) are the level-independent *QBD* and $M/G/1$ type *MC*s. These classes assume that the matrices of transition rates are block quasi-Toeplitz, i.e., the value of the block $\mathcal{Q}_{i,j}$ for $i > i^0,\ j \geq i-1$, depends on the difference $j-i$ but does not depend on i and j separately. The name quasi-Toeplitz goes back to the concept of the Toeplitz matrix. The prefix "quasi" reflects the possibility of a violation of the Toeplitz property of the generator for some low levels.

For $M/G/1$ type *MC*s, there exist matrices $\mathcal{Q}^-,\ \mathcal{Q}^0,\ \mathcal{Q}^{+k},\ k \geq 1$, such that

$$\mathcal{Q}_{i,i-1} = \mathcal{Q}^-,\ \mathcal{Q}_{i,i} = \mathcal{Q}^0,\ \mathcal{Q}_{i,i+k} = \mathcal{Q}^{+k},\ k \geq 1,\ i > i^0.$$

For level-independent *QBD*s, $\mathcal{Q}^{+k} = O$ for $k > 1$. For brevity, we will denote $\mathcal{Q}^{+1} = \mathcal{Q}^+$.

The classes of the level-independent *QBD* and $M/G/1$ type *MC*s were investigated in detail by M. Neuts; see, e.g., his seminal books [10,11]. Following M. Neuts, we assume that the matrix $\mathbf{Q} = \mathcal{Q}^- + \mathcal{Q}^0 + \sum\limits_{k=1}^{\infty} \mathcal{Q}^{+k}$ is irreducible.

The necessary and sufficient condition for the ergodicity of $M/G/1$ type *MC*s is given by M. Neuts in the following form.

Lemma 1. *$M/G/1$ type MC is ergodic if and only if the inequality*

$$\mathbf{q}\mathcal{Q}^-\mathbf{e} > \mathbf{q}\sum_{k=1}^{\infty} k\mathcal{Q}^{+k}\mathbf{e} \quad (2)$$

holds good where the row vector $\mathbf{q}$ *is the unique solution to the system*

$$\mathbf{qQ} = \mathbf{0},\ \mathbf{qe} = 1,$$

where $\mathbf{0}$ *is the zero row vector and the column vector* $\mathbf{e}$ *has all entries equal to 1.*

The following statement immediately follows from this lemma.

Corollary 1. *The level-independent QBD is ergodic if and only if the inequality*

$$\mathbf{q}\mathcal{Q}^-\mathbf{e} > \mathbf{q}\mathcal{Q}^+\mathbf{e} \quad (3)$$

holds good.

Sometimes, in application to a concrete queueing system, it is possible to analytically obtain the vector **q** and reduce inequalities (2) or (3) to a simple scalar form.

2.3. Level-Dependent Quasi-Birth-and-Death Processes and $M/G/1$ Type Markov Chains, Asymptotically Quasi-Toeplitz Markov Chains

The classes of the level-independent QBD and $M/G/1$ type MCs are extremely useful for the analysis of a variety of queueing models, in particular, various queueing models with an infinite buffer. However, the inherent feature of many important queueing systems is that the MC describing the behavior of the system is level-dependent. This implies that the blocks of the corresponding generator describing transition rates from the level i to the level j depend not only on the difference $j - i$ but also on i and j separately. It is mentioned in [12] that level-dependent QBDs are often more realistic and, while efficient and stable numerical solution techniques are available for level-independent QBDs, there are only a few approaches that try to exploit the block structure in the level-dependent case.

As the most important queueing models described by the level-dependent QBDs, retrial queueing models and queues with customer impatience have to be mentioned. More information about the retrial queues, real-world examples and known results can be found, e.g., in the books [13,14] and papers [15–19].

The importance of retrial queueing models stems from their suitability for modeling various real-world systems, including contact centers, delivery systems and the extremely popular wireless communication networks. The total intensity of retrials of customers staying in the orbit in the overwhelming majority of real systems and networks depends on the number of these customers. This makes the MC describing the behavior of the system level-dependent.

The impatience (reneging, abandonment, etc.) of customers also makes the M describing system behavior level-dependent.

Due to the practical importance of the analysis of level-dependent MCs, the notion of asymptotically quasi-Toeplitz Markov chains ($AQTMC$s) was introduced in the paper [20]. The rough intuitive definition of $AQTMC$s is as follows. The $AQTMC$ is an MC with the block upper-Hessenberg structure (1) of a generator that does not possess the quasi-Toeplitz property; however, in asymptotic ones, for very large values of the countable component of the chain, this property appears. A more exact and formal definition of $AQTMC$ is given below.

The main motivation for introducing $AQTMC$s is the necessity of considering retrial queues, with the retrial rate proportional to the number of customers in orbit, the $BMAP$ arrival process and the phase-type distribution of service times, see [21]. The significant difference between a system with an infinite buffer and a similar system with retrials is that in the former system, a new customer is immediately picked up for service from the buffer when some server is released. In an analogous situation in the latter system, there exists an interval during which the server (or servers) remains idle despite the customer's presence in the orbit. This period finishes via a new primary customer arrival or the retrial of a customer from the orbit. When the number of customers staying in the orbit infinitely increases, such a period becomes shorter and completely disappears in the limit. Therefore, the $AQTMC$ behaves in limit exactly as the corresponding limiting quasi-Toeplitz MC.

In the case when the generator of the $AQTMC$ is block tri-diagonal, the $AQTMC$ is a special case of the level-dependent QBD. The difference is that no assumptions about the blocks of a generator are made for the level-dependent QBD, while their asymptotic behavior is suggested for $AQTMC$.

It is worth noting that, due to the absence of a quasi-Toeplitz property of the MCs, which describes many queueing systems with customer retrials and (or) impatience, the problem of solving the infinite system of equilibrium equations for the stationary probabilities of the chain is quite difficult. Therefore, many researchers impose, from the early beginning, quite unrealistic assumptions about the considered system, like the orbit

capacity is finite, and the rate of retrials from the orbit is constant, independent of the current number of customers in the orbit, etc.

Some other researchers solve the infinite system of equilibrium equations via its truncation, see, e.g., [22]. The worst case, from a mathematical point of view, is when this is direct (brute force) truncation. Some equations are cut, and the remaining finite system is solved with the use of a computer. The better case is when the researchers use so-called soft truncation. One of the possible ways for soft truncation was offered by M.F. Neuts and B.M. Rao in [23]. This soft truncation suggests that the blocks of the generator for levels exceeding some fixed threshold become constant, independent of the level. In application to the analysis of a multi-server retrial queue, this means that after the number of customers in the orbit reaches some threshold (and until it drops below this threshold), the retrial rate becomes constant. But this suggestion is definitely not realistic because, usually, the total retrial rate is proportional to the number of customers staying in the orbit. When soft truncation is implemented, the results from [10] can be applied. If the generator is block tri-diagonal, i.e., the MC is the QBD, the vectors of the stationary probabilities of the states that belong to high levels have a matrix geometric form. When the truncation threshold is chosen suitably, the algorithm from [23] can give satisfactory results.

However, to apply this algorithm, it is necessary first to prove that, under the fixed set of transition rates, the considered QBD is ergodic and that the computed distribution is indeed the stationary distribution of the MC. Unfortunately, the paper [23] does not contain information about the conditions for ergodicity of the MC. Some researchers derive the ergodicity condition via the use of the results from [10] for level-independent QBD. But the obtained condition is the ergodicity condition for the MC with other dynamics of the MC when the current level of the MC is above the truncation threshold. Obviously, it is not the ergodicity condition for the initial level-dependent QBD.

The problem of the derivation of ergodicity conditions in the case of level-dependent MCs is very important but is not sufficiently addressed in the existing literature.

The condition given for the level-dependent QBDs in [24] is not a constructive one. It is given as a requirement for the convergence of some matrix series, the terms of which contain the infinite set of matrices (denoted as R_i in [24]) that are formally computed recursively. In the case of level-independent QBD, the recursion turns to the quadratic matrix equation. As it is known, see [10], the solution of this equation with the required properties exists only if the QBD is ergodic. Therefore, in the more complicated case of level-dependent QBD considered here, the situation is more difficult; the existence of a solution to the infinite recursion has to be justified, and at least the ergodicity of QBD has to be postulated. Thus, there is an evident vicious circle. To check the ergodicity, it is required to compute the matrices R_i which, in turn, may have a chance to be computed only if the QBD is ergodic.

2.4. Conditions for Ergodicity and Non-Ergodicity of Asymptotically Quasi-Toeplitz Markov Chains

Constructive sufficient conditions for the ergodicity and non-ergodicity of $AQTMC$s, a special case of which is an important class of level-dependent QBD, were presented in [20]. Here, we briefly reproduce the results relevant to our analysis from [20].

According to the definition of $AQTMC$s, an MC $\chi_t = \{i_t, n_t\}$ belongs to the class of $AQTMC$s if

(A) Its generator has the upper-Hessenberg structure (1);
(B) The following matrices Y_k, $k \geq 0$, exist:

$$Y_k = \lim_{i \to \infty} \mathcal{U}_i^{-1} \mathcal{Q}_{i,i+k-1} + \delta_{k,1} I, \; k \geq 0,$$

where $\delta_{k,1} = 1$ if $k = 1$ and $\delta_{k,1} = 0$; otherwise, $\mathcal{U}_i$ is the diagonal matrix with the diagonal entries defined by the moduli of the diagonal entries of the matrix $Q_{i,i}$. In other words,

$$\mathcal{U}_i = -I \circ \mathcal{Q}_{i,i},\ i \geq i^0,$$

where $\circ$ is the Hadamard product of matrices symbol, see, e.g., [25], and the matrix $\sum_{k=0}^{\infty} Y_k$ is the stochastic one;

(C) Some technical assumptions related to the requirement of the finiteness of the average size of the jump-up of the level of $AQTMC$ (see Theorem 4 in [20]) are fulfilled. These assumptions are evidently implemented, e.g., under the suggestion that $Y_k = O$ for $k > K+1$ where K is a finite integer, $K \geq 1$. Thus, below, we impose this suggestion.

A sufficient condition for the ergodicity of $AQTMC$s proven in [20] is given as follows.

Let us introduce the matrix generating function $Y(z) = \sum_{k=0}^{\infty} Y_k z^k,\ |z| < 1$. Ergodicity conditions are different depending on the irreducibility or reducibility of the matrix $Y(1)$. It is worth noting that although we supposed above that the matrix $\mathbf{Q} = \mathcal{Q}^- + \mathcal{Q}^0 + \sum_{k=1}^{\infty} \mathcal{Q}^{+k}$ is irreducible, the matrix $Y(1)$ can be (and often is) reducible. Therefore, two variants of the ergodicity condition have to be analysed.

Lemma 2. *If the matrix $Y(1)$ is irreducible, the sufficient condition for the ergodicity of $AQTMC$s is the fulfillment of the inequality*

$$\mathbf{y} \sum_{k=1}^{\infty} k Y_k \mathbf{e} < 1 \tag{4}$$

where the row vector $\mathbf{y}$ is the unique solution of the system

$$\mathbf{y} = \mathbf{y} \sum_{k=0}^{\infty} Y_k,\ \ \mathbf{y}\mathbf{e} = 1. \tag{5}$$

If the matrix $Y(1)$ is reducible, then, by means of the coordinated permutation of rows and columns, the matrix $Y(1)$ can be represented in the normal form, see [26],

$$Y(1) = \begin{pmatrix} Y^{\{1,1\}} & O & O & \dots & O & O & O & O \\ O & Y^{\{2,2\}} & O & \dots & O & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ O & O & O & \dots & Y^{\{m,m\}} & O & O & O \\ Y_{m+1,1} & Y_{m+1,2} & Y_{m+1,3} & \dots & Y_{m+1,m} & Y^{\{m+1,m+1\}} & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ Y_{s,1} & Y_{s,2} & Y_{s,3} & \dots & Y_{s,m} & Y_{s,m+1} & \dots & Y^{\{s,s\}} \end{pmatrix}$$

where $Y^{\{l,l\}},\ l = 1, \dots, m$ are irreducible stochastic matrices, the matrices $Y^{\{l,l\}},\ l = m+1, \dots, s$, are irreducible matrices, and for each $l,\ l = m+1, \dots, s$, at least one of the matrices $Y_{l,1}, \dots, Y_{l,l-1}$ is non-zero.

Correspondingly to the normal form of the matrix $Y(1) = \sum_{k=0}^{\infty} Y_k$, all matrices $Y_k,\ k \geq 0$, can also be represented in a similar form. In particular, we denote by $Y_k^{\{l,l\}}$ the diagonal blocks of the matrix $Y_k,\ k \geq 0,\ l = 1, \dots, m$.

According to [20], the following statement is true.

Lemma 3. *In the case of the reducible matrix $Y(1)$, the sufficient condition for the ergodicity of $AQTMC$s is the fulfillment of all m inequalities*

$$\mathbf{y}^{\{l\}} \sum_{k=1}^{\infty} k Y_k^{\{l,l\}} \mathbf{e} < 1,\ l = 1, \dots, m, \tag{6}$$

where the row vector $\mathbf{y}^{\{l\}}$ *is the unique solution of the system*

$$\mathbf{y}^{\{l\}} = \mathbf{y}^{\{l\}} \sum_{k=0}^{\infty} Y_k^{\{l,l\}}, \ \ \mathbf{y}^{\{l\}}\mathbf{e} = 1, \ l = 1, \ldots, m. \tag{7}$$

The use of Lemmas 2 and 3 allows for the determination of ergodicity conditions for various queueing systems. Sometimes, these conditions can be easily verified numerically. Finite systems (5) or (7) of the linear algebraic equations are solved, and their solutions are substituted into inequalities (6) or (7). Sometimes, the inequalities can be reduced to a nice scalar form.

However, the application of these conditions to the analysis of concrete queueing models requires preliminary verification that the *MC* describing a queueing model indeed belongs to the class of *AQTMC*s. To this end, a computation of the blocks $Y_k, \ k \geq 1$, of the one-step transition probability matrix for the limiting discrete-time jump *MC* is necessary. It may not be easy.

It appears that if the customers arriving at the system can balk (abandon) the system with the probability tending to 1 when queue length upon arrival infinitely increases, or if the customers waiting in the system are impatient or non-persistent, sometimes, this verification and the other steps for the proof that the considered *MC* is ergodic are not necessary. It can be stated that the considered *MC* belongs to the class of *AQTMC*s and is ergodic (the corresponding queueing system has a stationary regime of operation) for any set of the system parameters.

Here, we present the results, which allow us to skip, under quite non-restrictive assumptions, the necessity of proving the affiliation of the considered *MC* to the class of *AQTMC*s, including the computation of the matrices $Y_k, \ k \geq 0$. This justifies the direct use of the algorithms developed for the computation of the stationary distribution of *AQTMC*s in [20,27–29] to compute the stationary distribution of the considered queueing system. The use of the algorithms from [20,27] requires certain analytical derivations (calculation of the limits of some matrices) to obtain the explicit form of the blocks Y_k of the one-step transition probability matrix of the limiting discrete-time *MC* for the *AQTMC*. The algorithms proposed in [28,29] do not need such derivations because they operate directly only with the blocks of the generator of the *AQTMC*. It should also be stressed that the results of this paper render unnecessary the derivation and control of the fulfillment of an ergodicity condition for *MC*s describing various queueing systems because it is shown here that these *MC*s are always ergodic due to customers' balking, impatience or non-persistence.

3. Impact of Customers' Impatience in the Systems Described, in the Absence of Impatience, by the Level-Independent *QBD* and *M*/*G*/1 Type *MC*

3.1. Problem Statement

Let us consider a queueing system, the behavior of which, in the absence of customer impatience and non-persistence, is described by a regular irreducible continuous-time *MC* $\tilde{\zeta}_t = \{i_t, n_t\}$, where the countable component i_t defines the number of customers in the system, $i_t \geq 0$. The process n_t describes the transitions of a finite component, defining, along with the component i_t,, the dynamics of the system.

Let us now assume that the customers who arrive to receive service in the system are impatient. Impatience means the ability of a customer to depart (renege) from the system while waiting in a buffer. We denote the *MC* describing this system as $\zeta_t = \{i_t, n_t\}$. Note that the *MC*s $\tilde{\zeta}_t = \{i_t, n_t\}$ and $\zeta_t = \{i_t, n_t\}$ have the same state space but different dynamics.

We assume that if the current state of the queueing system belongs to level i, then, during an interval of very small length Δ, with the probability $\alpha_i\Delta + o(\Delta)$, one customer permanently departs from the system, and the finite component n_t makes the transition possible. The matrix of the corresponding transition probabilities is denoted by Ψ. It is most likely that $\Psi = I$. However, for generality, we admit any stochastic matrix Ψ. We suppose

that the impatience rates α_i tend to infinity when i goes to infinity. Note that the most popular and reasonable dependence of α_i on i is $\alpha_i = i\alpha$ or $\alpha_i = \max\{0, i - J\}\alpha$. Here, α is the impatience rate (the parameter of exponentially distributed patience time) of individual customers, and it is assumed that customers renege from the system due to impatience independently of each other; J is the number of customers that cannot renege from the system. For example, if the component i_t is the number of customers in the system, and only customers staying in the buffer can renege, then J corresponds to the current number of customers that receive service.

Our aims are to prove that the *MC* ζ_t belongs to the class of $AQTMC$s and is ergodic for any values of the system parameters.

3.2. Problem Solution

Firstly, we consider the partial case when the *MC* $\tilde{\zeta}_t = \{i_t, n_t\}$ describing the dynamic of the system with *patient* customers is the level-independent QBD. This means that the generator $\tilde{\mathcal{Q}}$ of this *MC* is the block tri-diagonal matrix having the blocks $\tilde{\mathcal{Q}}_{i,j}$, of the form:

$$\tilde{\mathcal{Q}}_{i,i} = \mathcal{Q}^0,\ \tilde{\mathcal{Q}}_{i,i+1} = \mathcal{Q}^+,\ \tilde{\mathcal{Q}}_{i,i-1} = \mathcal{Q}^-,\ i > i^0.$$

It is easy to see that the *MC* $\zeta_t = \{i_t, n_t\}$ describing the dynamic of the system with *impatient* customers is the level-dependent QBD. The generator $\mathcal{Q}$ of this *MC* is the block tri-diagonal matrix having the blocks $\mathcal{Q}_{i,j}$, which define transition rates between the states that belong to the level i and states that belong to the level j, of the form:

$$\mathcal{Q}_{i,i} = \mathcal{Q}^0 - \alpha_i I_{N+1},\ \mathcal{Q}_{i,i+1} = \mathcal{Q}^+,\ \mathcal{Q}_{i,i-1} = \mathcal{Q}^- + \alpha_i \Psi,\ i > i^0.$$

Lemma 4. *The level-dependent QBD* ζ_t *belongs to the class of AQTMCs.*

Proof. It is clear that the block tri-diagonal structure of the generator $\mathcal{Q}$ of the *MC* ζ_t is the special case of the structure (1).

It is evident that the matrix $\mathcal{U}_i$ for the *MC* ζ_t is defined by

$$\mathcal{U}_i = \alpha_i I_{N+1} - \hat{\mathcal{Q}}^0,$$

where $\hat{\mathcal{Q}}^0$ is the diagonal matrix with the diagonal entries defined by the diagonal entries of the matrix $\mathcal{Q}^0$.

Calculating the limiting matrices Y_k in the definition of $AQTMC$, it is easy to see that these limits indeed exist and are defined by

$$Y_0 = \Psi,\ Y_1 = O,\ Y_2 = O.$$

Thus, the *MC* ζ_t satisfies the definition of $AQTMC$s. The lemma is proven. □

Theorem 1. *The MC* ζ_t *is ergodic for any choice of the system parameters.*

Proof. To prove the theorem, we apply Lemma 2 or Lemma 3. If the matrix Ψ is irreducible, then the ergodicity condition (4) turns to the inequality $0 < 1$, which is always true.

If the matrix Ψ is reducible, then the ergodicity conditions (6) with an account of evident equalities $Y_1^{\{l,l\}} = Y_2^{\{l,l\}} = O,\ l = 1, \ldots, m$, also turn to inequality $0 < 1$, which is always true.

Thus, we have proven that the *MC* ζ_t is ergodic for any choice of the system parameters. □

Now, let the *MC* $\tilde{\zeta}_t^* = \{i_t, n_t\}$ not be the level-independent QBD but the more general $M/G/1$ type *MC* having the generator, which is a particular case of the generator of form (1).

Let us assume that for $i > i^0$, the blocks of the generator $\tilde{Q}$ of this chain are defined by the formulas

$$\tilde{Q}_{i,i} = Q^0,\ \tilde{Q}_{i,i-1} = Q^-,\ \tilde{Q}_{i,i+k} = Q^{+k},\ k = 1, \dots, K,\ \tilde{Q}_{i,i+k} = O,\ k > K.$$

Here, K is an integer number, $K \geq 1$. When $K = 1$, the *MC* $\tilde{\zeta}_t^*$ is the *QBD*. The case of *QBD* has been analyzed above. Now, let us assume that $K > 1$. Note that the requirement that the number K be finite is not essential.

It is easy to see that the blocks of the generator of the *MC* $\zeta_t^* = \{i_t, n_t\}$ describing the corresponding system with customer impatience are defined by

$$Q_{i,i} = Q^0 - \alpha_i I_{N+1},\ Q_{i,i+k} = Q^{+k},\ k = 1, \dots, K,\ Q_{i,i-1} = Q^- + \alpha_i \Psi,\ i > i^0.$$

Lemma 5. *The MC* ζ_t^* *belongs to the class of AQTMCs.*

The proof of this lemma repeats the proof of Lemma 4. However, here, not just two matrices Y_1 and Y_2 in the definition of *AQTMC* but all matrices $Y_k,\ k = 1, 2, \dots, K+1$, are equal to O.

Theorem 2. *The MC* ζ_t^* *is ergodic for any choice of the system parameters.*

The proof of this theorem repeats the proof of Theorem 1, taking into account that, here, all matrices $Y_k,\ k = 1, \dots, K+1$, are equal to O, while the matrix Y_0 is stochastic.

3.3. The Case of the Impatience Rate Dependent on the Value of Both Components of the Chain

In the problem statement, we have assumed that if the current state of the queueing system belongs to level i, then, during an interval of very small length Δ, with the probability $\alpha_i \Delta + o(\Delta)$, one customer permanently departs from the system, and the impatience rate α_i tends to infinity when i tends to infinity.

Analyzing the proof of the ergodicity of the considered *MC* for all values of the system parameters, it is not difficult to see that the obtained result can be generalized as follows.

Let us assume that if the state of the *MC* ζ_t^* is (i, n), then the probability that one customer permanently departs from the system during an interval of a very small length Δ is equal to $\alpha_i^{(n)} \Delta + o(\Delta)$ and

$$\alpha_i^{(n)} \to \infty \text{ for all values of } n = 0, 1, \dots, N.$$

As above, the component n_t can make transitions defined by the stochastic matrix Ψ at the moment of the customer reneging.

In other words, here, we consider a more general case than the one considered above by assuming the possibility of having different impatience rates under different states of the finite component of the *MC*.

It is easy to prove, by analogy with the previous statements, the following assertions.

Lemma 6. *The MC* ζ_t^* *belongs to the class of AQTMCs.*

Theorem 3. *If the impatience rates* $\alpha_i^{(n)}$ *tend to infinite when i approaches infinite for any value of* $n,\ n = 0, 1, \dots, N$, *then the MC* ζ_t^* *and its particular case* ζ_t *are ergodic for any choice of the system parameters.*

The proof repeats the proof of Theorem 1 because it is easy to check that the matrix Y_0 is equal to the stochastic matrix Ψ while all other limiting matrices $Y_k,\ k \geq 1$, are equal to zero.

3.4. The System with Customers Balking

Let us return to the queueing system described by the *MC* $\tilde{\zeta}_t = \{i_t, n_t\}$ defined in Section 2.1. It is quite typical for many real-world systems that the queue length is visible for arriving customers, and they can join or abandon the queue with a probability depending on the current queue length. In this subsection, we briefly consider such a scenario when the customers can abandon (balk) the system upon arrival. Namely, let us assume that the customer (or a batch of customers) arriving to the system when the current number of customers in the queue is equal to i (i.e., the *MC* $\tilde{\zeta}_t = \{i_t, n_t\}$ resides in the state belonging to the level i), joins the queue with the probability q_i and abandons the system with the complimentary probability.

We omit consideration of the simpler case when $K = 1$ and assume that $K \geq 1$. Let us denote the *MC* describing the queue with customers balking by η_t^*. It is easy to verify that the blocks of the generator of the *MC* $\eta_t^* = \{i_t, n_t\}$ are defined via the blocks of the *MC* $\tilde{\zeta}_t$ as follows:

$$\mathcal{Q}_{i,i} = \mathcal{Q}^0 + (1 - q_i) \sum_{k=1}^{K} \mathcal{Q}^{+k},\ \mathcal{Q}_{i,i+k} = q_i \mathcal{Q}^{+k},\ k = 1, \ldots, K,\ \mathcal{Q}_{i,i-1} = \mathcal{Q}^-,\ i > i^0.$$

Lemma 7. *The MC* η_t^* *belongs to the class of AQTMCs.*

Proof. It is easy to see that, here, the matrix $\mathcal{U}_i$ is defined by

$$\mathcal{U}_i = -I \circ (\mathcal{Q}^0 + (1 - q_i) \sum_{k=1}^{K} \mathcal{Q}^{+k}),\ i > 0,$$

and the limiting matrices Y_k exist and are defined by formulas

$$Y_0 = -(I \circ \mathcal{T})^{-1} \mathcal{Q}^-,\ Y_1 = I - (I \circ \mathcal{T})^{-1} \mathcal{T},\ Y_k = O,\ k = 2, \ldots, K,$$

where

$$\mathcal{T} = \mathcal{Q}^0 + \sum_{k=1}^{K} \mathcal{Q}^{+k}.$$

This implies that the *MC* η_t^* indeed belongs to the class of *AQTMC*. □

Theorem 4. *If the probabilities* q_i *tend to zero when i approaches infinity, then the MC* η_t^* *is ergodic for any choice of the system parameters.*

Proof. We obtain that the stochastic matrix $Y(1) = \sum_{k=0}^{K} Y_k$ is, here, the sum of only two sub-stochastic matrices, Y_0 and Y_1, and the matrix Y_0 is a non-zero matrix. Taking into account the explicit form of the matrices Y_0 and Y_1, we obtain that the matrix $Y(1)$ is defined by the formula

$$Y(1) = I - (I \circ \mathcal{T})^{-1}(\mathcal{Q}^- + \mathcal{T}) = I - (I \circ \mathcal{T})^{-1}\mathbf{Q}.$$

It was supposed above that the matrix $\mathbf{Q}$ is irreducible. A multiplication of this matrix from the left by the diagonal matrix with all non-zero diagonal entries and summing up with the identity matrix cannot make the resulting matrix reducible. Thus, the matrix $Y(1)$ is irreducible.

Therefore, inequality (4) as a sufficient condition for the ergodicity of *AQTMC*s can be rewritten here in the form $\mathbf{y}Y_0\mathbf{e} > \mathbf{y}Y_2\mathbf{e}$ and always holds good because the matrix Y_0 is non-zero sub-stochastic, while the matrix Y_2 is a zero matrix. The theorem is proven. □

Remark 1. *The obtained result holds good also in a more general situation when the joining probabilities have the form* $q_i^{(k)}$*, i.e., they depend on both the queue length i at the arrival moment and the number k of customers in the arrived batch of customers. The MC* ζ_t^* *and its particular case*

ζ_t are ergodic for any choice of the system parameters if the probabilities $q_i^{(k)}$ tend to zero for all values of k, $k = 1, 2, \ldots, K$, when i approaches infinity.

4. Impact of Customers' Impatience in the Multi-Server Retrial Queueing Systems

4.1. Problem Statement

Let us consider the N-server retrial queueing system with patient customers. A customer who arrives at the system when less than N servers are busy immediately starts service. A customer who arrives when all servers are busy moves to the special virtual place called orbit and retries to enter service after exponentially distributed times. Let the total retrial rate from the orbit be equal to ν_i when the number of customers staying in the orbit of an infinite capacity is equal to i, $i \geq 1$. We assume that $\nu_0 = 0$ and that there exists a finite or infinite limit of values ν_i when i tends to infinity. The most popular in the literature dependencies of ν_i on i are the classical retrial policy, when $\nu_i = i\nu$, where ν is an individual retrial rate, and the constant retrial rate when ν_i is equal to the constant independent of i.

If some server is available at a retrial moment, the retrying customer occupies the server and departs from the system after service completion. If all servers are busy at the retrial moment, the retrying customer returns to the orbit.

The behavior of the system is described by the three-dimensional MC $\tilde{\xi}_t = \{i_t, n_t, r_t\}$, where i_t is the number of customers in the orbit, $i_t \geq 0$, n_t is the number of busy servers, $n_t = 0, 1, \ldots, N$, r_t is the state of some auxiliary components of the chain, $r_t = 1, \ldots, R$, at the moment t. As in the previous section, we start analysis from the case when the generator $\tilde{\mathcal{Q}}$ of the MC $\tilde{\xi}_t$ has the block tri-diagonal structure with the non-zero blocks $\tilde{\mathcal{Q}}_{i,j}$ defined by the formulas

$$\tilde{\mathcal{Q}}_{i,i} = \mathcal{A} - \nu_i \mathcal{I}^0,\ i \geq 0, \tilde{\mathcal{Q}}_{i,i+1} = \mathcal{C},\ i \geq 0,\ \tilde{\mathcal{Q}}_{i,i-1} = \nu_i \mathcal{B},\ i \geq 1.$$

Here, the subgenerator $\mathcal{A}$ defines the transition rates of the components $\{n_t, r_t\}$ of the MC $\tilde{\xi}_t$ that do not lead to a change in the value of the component i_t of this chain. The matrix $\mathcal{C}$ defines the rates of transitions of the components $\{n_t, r_t\}$ of the MC $\tilde{\xi}_t$ that imply the increase in the value of the component i_t by 1. The matrix $\mathcal{B}$ defines the transition probabilities of the components $\{n_t, r_t\}$ of the MC $\tilde{\xi}_t$ that imply the decrease in the value of the component i_t by 1. The matrix $\mathcal{I}^0$ is defined by the formula $\mathcal{I}^0 = \tilde{I} \otimes I_R$, where $\tilde{I} = \mathrm{diag}\{1, 1, \ldots, 1, 0\}$, $\mathrm{diag}\{\ldots\}$ denotes the diagonal matrix with the diagonal entries listed in the brackets, and $\otimes$ is the symbol of Kronecker product of matrices, see [30].

In detail, the matrix $\mathcal{C}$ is defined by the formula $\mathcal{C} = (I - \tilde{I}) \otimes C$ where the matrix C defines the transition rates of the component r_t of the MC $\tilde{\xi}_t$ at the moment of a customer's arrival to the system in the presence of N busy servers. The matrix $\mathcal{B}$ is defined by the formula $\mathcal{B} = \mathrm{diag}^+\{B_0, B_1, \ldots, B_{N-1}\}$ where the $\mathrm{diag}^+\{\ldots\}$ denotes the matrix having all zero blocks except the blocks B_n above the diagonal, $n = 0, 1, \ldots, N-1$. The block B_n defines the transition probabilities of the component r_t of the MC $\tilde{\xi}_t$ at the moment of a retrying customer service beginning in the presence of n busy servers.

The concrete form of the described blocks for the particular case of the considered general retrial queueing system, such as the $BMAP/PH/N$ type retrial queue, can be found in [21,31]. Here, $BMAP$ denotes the batch Markov arrival process; for details, see, e.g., [32–34]. The PH denotes the phase-type distribution of service times. Details about the PH distribution can be found, e.g., in [10,32,34,35].

In [21], the component r_t of the MC $\tilde{\xi}_t$ is the set of components such as the underlying process of the $BMAP$ flow of customers and the underlying processes of the PH distribution of service time on all busy servers. In [31], the component r_t of the MC $\tilde{\xi}_t$ is the set of components such as the underlying process of the $BMAP$ flow of customers and the number of servers providing service at each phase. For more details about these two different ways for tracking the PH distribution of service time in busy servers, see, e.g., [36]. For the purposes of this paper, the explicit form of the blocks $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ does not matter.

Let us now modify this queueing system by assuming that the customers staying in the orbit are impatient. The mechanism of orbiting customers reneging from the orbit is the same as the mechanism of waiting customers reneging from the buffer described in the previous section. The total rate of customers reneging from the orbit when the number of customers in the orbit is equal to i, $i \geq 1$, is denoted by α_i. We suppose that $\lim\limits_{i\to\infty} \alpha_i = \infty$. We denote the matrix of transition probabilities of the components $\{n_t, r_t\}$ at the moment of a customer reneging as Ψ. Note that, in this section, in which we separate two finite components, the size of the square matrix Ψ is $(N+1)R$. In the previous section, the corresponding matrix had a size $(N+1)$.

The *MC* describing the considered retrial queue in the presence of customer impatience is denoted by ξ_t .

Our aim is to prove that the *MC* ξ_t is affiliated to the class of *AQTMC* and is ergodic for any value of the system parameters.

4.2. Problem Solution

As in the previous section, we start our analysis from the case when the generator $\mathcal{Q}$ of the *MC* ξ_t has the block tri-diagonal structure.

It can be verified that the non-zero blocks $\mathcal{Q}_{i,j}$ of this generator are defined by the formulas

$$\mathcal{Q}_{i,i} = \mathcal{A} - \nu_i \mathcal{I}^0 - \alpha_i I_{(N+1)R},\ i \geq 0, \mathcal{Q}_{i,i+1} = \mathcal{C},\ i \geq 0, \mathcal{Q}_{i,i-1} = \nu_i \mathcal{B} + \alpha_i \Psi, i \geq 1. \quad (8)$$

Lemma 8. *The MC ξ_t belongs to the class of AQTMCs.*

Proof. Let us denote $\hat{\mathcal{A}}$ as the diagonal matrix with the diagonal entries defined by the moduli of the diagonal entries of the matrix $\mathcal{A}$. Also denoted as $\mathcal{U}_i$ is the matrix $\mathcal{U}_i = \hat{\mathcal{A}} + ((\nu_i + \alpha_i)\tilde{I} + \alpha_i(I - \tilde{I})) \otimes I_R$.

It is not difficult to verify that the following limits exist:

$$Y_k = \lim_{i\to\infty} \mathcal{U}_i^{-1} \mathcal{Q}_{i,i+k-1} + \delta_{k,1} I,\ k = 0, 1, 2,$$

and are defined as

$$Y_1 = Y_2 = O,\ Y_0 = \frac{\gamma}{\gamma+1}(\tilde{I} \otimes I_R)\mathcal{B} + (\frac{1}{\gamma+1}\tilde{I} \otimes I_R + (I - \tilde{I}) \otimes I_R)\Psi. \quad (9)$$

Here,

$$\gamma = \lim_{i\to\infty} \frac{\nu_i}{\alpha_i}.$$

If $\gamma = 0$, i.e., α_i tends to infinity more fast than ν_i, then, $Y_0 = \Psi$. If $\gamma = \infty$, i.e., α_i tends to infinity more slowly than ν_i, then, $Y_0 = (\tilde{I} \otimes I_R)\mathcal{B} + ((I - \tilde{I}) \otimes I_R)\Psi$. If γ is a finite positive number, the matrix Y_0 is a stochastic matrix defined in Formula (9). □

Therefore, results from [20] can be used for the derivation of a sufficient condition for the ergodicity of the *MC* ξ_t. Note that when $K = 1$, inequalities (4) and (6) in the ergodicity conditions for *AQTMC*s take the form

$$\mathbf{y} Y_0 \mathbf{e} > \mathbf{y} Y_2 \mathbf{e}, \quad \mathbf{y}^{\{l\}} Y_0^{\{l,l\}} \mathbf{e} > \mathbf{y}^{\{l\}} Y_2^{\{l,l\}} \mathbf{e},\ \ l = 1, \ldots, m.$$

Because, according to (9), the matrix Y_0 is a stochastic one while $Y_2 = O$, these inequalities are trivially fulfilled.

Therefore, the following statement is proven.

Theorem 5. *The MC ξ_t describing the retrial queueing model with impatient customers is ergodic for any choice of the system parameters.*

Let ξ_t^* be the *MC* which is the generalization of the *MC* ξ_t to the case when the generator of the chain has not three but $K+1$ non-zero block diagonals, similar to the *MC* ζ_t^* considered in the previous section as the generalization of the *MC* ζ_t.

The validity of the following corollaries is easily established.

Corollary 2. *The MC ξ_t^* is affiliated to the class of AQTMCs.*

Corollary 3. *The MC ξ_t^* is ergodic for any choice of the system parameters.*

Remark 2. *If arriving customer balking (refusal to go to orbit upon arrival if all servers are busy) would be incorporated into the considered retrial queueing system described by the MC $\tilde{\xi}_t$, and the probability of moving to orbit q_i tends to zero, then, the modified retrial queueing model will be stable for all values of the system parameters.*

5. Impact of Customers Non-Persistence in the Multi-Server Retrial Queueing Systems

5.1. Problem Statement

Let us consider the same three-dimensional *MC* $\tilde{\xi}_t = \{i_t, n_t, r_t\}$ describing the behavior of the multi-server retrial queueing system as in the previous section. The generator $\tilde{\mathcal{Q}}$ of this chain is defined by the blocks $\tilde{\mathcal{Q}}_{i,j}$ given by Formula (8).

Let us now modify the queueing system by assuming that if a customer makes the retrial in the presence of i customers in the orbit, and all servers are busy, then, with the probability $p_i,\ i \geq 1$,, this customer returns to the orbit, and with the complementary probability, it departs from the system permanently. We assume that there is a limit

$$p = \lim_{i \to \infty} p_i$$

and $p < 1$.

This assumption is fulfilled, in particular, when the customers staying in the orbit decide whether to return to the orbit in the case when all servers are busy at the retrial moment independently of each other with the probability $p,\ p < 1$.

Notions of customer impatience and non-persistence are close. Both impatience and non-persistence cause the same effect, namely, that the customer permanently leaves the system. The difference is that the departure due to impatience can occur at an *arbitrary* moment during the customer staying in the orbit, while the departure due to non-persistence can occur at any *retrial* moment. In the context of retrial queues, it is more reasonable for the customer waiting in the orbit to check the status of the servers before departing, whether or not all servers are busy at this moment. However, sometimes, the customers can leave the system without making the "last" trial. For example, in modeling the cell of the mobile communication network, the retrying customer can decide to stop retrials after any unsuccessful retrial and can also leave the cell due to the loss of connection to the base station under the move to another cell.

Let us denote by $\hat{\xi}_t$ the *MC* describing the dynamics of the system with non-persistent customers. Additionally, let us assume that a customer departure from the system due to non-persistence does not cause any changes in the value of the components $\{n_t, r_t\}$ of the *MC* $\hat{\xi}_t$. The contrary case can be considered analogously.

The aim of this section is to briefly show that the *MC* $\hat{\xi}_t$ is always ergodic.

5.2. Problem Solution

As in the previous sections, we start analysis from the case when the generator $\mathcal{Q}$ of the *MC* $\hat{\xi}_t$ has the block tri-diagonal structure.

It can be verified that the non-zero blocks $\mathcal{Q}_{i,j}$ of this generator are defined by formulas

$$\mathcal{Q}_{i,i} = \mathcal{A} - \nu_i \hat{\mathcal{I}}_i^0,\ i > 0, \mathcal{Q}_{i,i+1} = \mathcal{C},\ i > 0, \mathcal{Q}_{i,i-1} = \nu_i \hat{\mathcal{B}}_i, i > 1,$$

where

$$\hat{\mathcal{I}}_i^0 = (\tilde{I} + (1-p_i)(I-\tilde{I})) \otimes I_R,\ i \geq 1,$$
$$\hat{\mathcal{B}}_i = \mathcal{B} + (1-p_i)(I-\tilde{I}) \otimes I_R,\ i > 1.$$

Lemma 9. *The MC $\hat{\xi}_t$ belongs to the class of AQTMCs.*

Proof. Let, as in the previous section, $\hat{\mathcal{A}}$ be the diagonal matrix with the diagonal entries defined by the moduli of the diagonal entries of the matrix $\mathcal{A}$.

It is easy to see that the matrix $\mathcal{U}_i$ appearing in the definition of the *AQTMC* is defined by the formula

$$\mathcal{U}_i = \hat{\mathcal{A}} + \nu_i \hat{\mathcal{I}}_i^0.$$

It is not difficult to verify that the following limits

$$Y_k = \lim_{i\to\infty} \mathcal{U}_i^{-1} \mathcal{Q}_{i,i+k-1} + \delta_{k,1} I,\ k = 0,1,2,$$

indeed exist and are defined as

$$Y_1 = Y_2 = O,\ Y_0 = \mathcal{B} + (I-\tilde{I}) \otimes I_R.$$

This completes the proof of the lemma. □

Again, the matrix Y_0 is a stochastic one, and, irrespective of whether the matrix Y_0 is irreducible or reducible, the fact that $Y_1 = Y_2 = O$ implies the trivial fulfillment of inequalities (4) or (6) in a sufficient condition of ergodicity of the *MC*.

Therefore, the following statement is proven.

Theorem 6. *The MC $\hat{\xi}_t$ describing the retrial queueing model with non-persistent customers is ergodic for any choice of the system parameters.*

Let $\hat{\xi}_t^*$ be the *MC* which is the generalization of the *MC* $\hat{\xi}_t$ to the case when the generator of the chain has not three but $K+1$ non-zero block diagonals, similar to the *MC* ζ_t^* considered in Section 2 as the generalization of the *MC* ζ_t.

The validity of the following corollary is easily established.

Corollary 4. *The MC $\hat{\xi}_t^*$ is affiliated to the class of AQTMCs.*

Corollary 5. *The MC $\hat{\xi}_t^*$ is ergodic for any choice of the system parameters.*

Remark 3. *Theorem 5 states that the retrial queue is always stable (the corresponding MC is ergodic) if the customers are impatient. Theorem 6 states that the retrial queue is always stable if the customers are non-persistent. It is easy to check that the retrial queue is always stable if the customers are both impatient and non-persistent.*

For reader convenience, the random processes used and analyzed in the text are summarised in Table A1 in Appendix A.

6. Conclusions

In this paper, it is shown that for a great variety of queueing models, including a wide range of variants of the *BMAP*/*PH*/*N*-type queues, with an infinite buffer or orbit, the following is true. If the customers waiting in the buffer or the orbit are impatient with an infinitely increasing total impatience rate, then, the multidimensional *MC* describing the behavior of the system is affiliated to the class of *AQTMC*s and is always ergodic. The same is valid for systems with customers balking upon arrival as well as for the retrial queue with non-persistent customers, with the limiting value of the probability to return to the orbit when all servers are busy at the retrial moment less than 1.

The account of customer impatience, balking or non-persistence usually leads, generally speaking, to an essential complication in the computation of the stationary distribution compared to the systems with patient customers due to the space-inhomogeneous behavior of the corresponding MC. But, as it is shown in this paper, the obtained MC belongs to the class of $AQTMC$s. This has the following two important implications : (i) the question about the existence of a stationary distribution (under non-restrictive assumptions) always has a positive answer; (ii) effective and numerically stable algorithms from [20,27–29] can be used to compute the stationary distribution of the considered queueing system instead of various truncation schemes popular in the existing literature.

The presented results are planned to be extended to the case when the impatience rate depends not only on the value of the denumerable component but also on the value of the finite component of the MC and tends to infinity not mandatory for *all* but at least for some values of the finite component.

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

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

Appendix A

Table A1. Processes denotation.

Process Denotation	System	Specifics	Section
$\tilde{\zeta}_t$	QBD	NO	Section 2.2
ζ_t	QBD	IMP	Section 2.2
$\tilde{\zeta}_t^*$	$M/G/1$	NO	Section 2.2
ζ_t^*	$M/G/1$	IMP	Sections 2.2 and 2.3
η_t^*	$M/G/1$	BALK	Section 2.4
$\tilde{\xi}_t$	QBD retrial	NO	Section 3.1
ξ_t	QBD retrial	IMP	Section 3.1
ξ_t^*	$M/G/1$ retrial	IMP	Section 3.1
$\hat{\xi}_t$	QBD retrial	NONPER	Section 4.1
$\hat{\xi}_t^*$	$M/G/1$ retrial	NONPER	Section 4.1

The first column of the table presents a process denotation. The second column defines the type of the system (or the MC) described by this process. QBD means the Quasi-Birth-and-Death process, and $M/G/1$ denotes the $M/G/1$ type MC. These symbols supplemented by the word "retrial" correspond to the systems of the same type but with the retrials. The third column identifies the specifics of the system described by the corresponding process. Here, IMP means that the customers are impatient, and BALK means that the customers can balk upon arrival. NONPER means that the retrial customers are non-persistent. NO means that the system does not have any specifics like denoted by the symbols IMP, BALK and NONPER.

References

1. Palm, C. Methods of judging the annoyance caused by congestion. *TELE* **1953**, *4*, 189–208.
2. Barrer, D.Y. Queuing with impatient customers and ordered service. *Oper. Res.* **1957**, *5*, 650–656. [CrossRef]

3. Haight, F.A. Queueing with reneging. *Metr. Int. J. Theor. Appl. Stat.* **1959**, *2*, 186–197. [CrossRef]
4. Sasieni, M.W. Double queues and impatient customers with an application to inventory theory. *Oper. Res.* **1961**, *9*, 771–781. [CrossRef]
5. Rao, S.S. Queuing with balking and reneging in $M/G/1$ systems. *Metr. Int. J. Theor. Appl. Stat.* **1967**, *12*, 173–188.
6. De Kok, A.G.; Tijms, H.C. A queueing system with impatient customers. *J. Appl. Probab.* **1985**, *22*, 688–696. [CrossRef]
7. Wang, K.; Li, N.; Jiang, Z. Queueing system with impatient customers: A review. In Proceedings of the 2010 IEEE International Conference on Service Operations and Logistics, and Informatics, Qingdao, China, 15–17 July 2010; pp. 82–87.
8. Sharma, S.; Kumar, R.; Soodan, B.S.; Singh P. Queuing models with customers' impatience: A survey. *Int. J. Math. Oper. Res.* **2023**, *26*, 523–547. [CrossRef]
9. Stanford, R.E. On queues with impatience. *Adv. Appl. Probab.* **1990**, *22*, 768–769. [CrossRef]
10. Neuts, M. *Matrix-Geometric Solutions in Stochastic Models*; The Johns Hopkins University Press: Baltimore, MD, USA, 1981.
11. Neuts, M. *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*; Marcel Dekker: New York, NY, USA, 1989.
12. Baumann, H.; Sandmann, W. Numerical solution of level dependent quasi-birth-and-death processes. *Procedia Comput. Sci.* **2010**, *1*, 1561–1569. [CrossRef]
13. Falin, G.I.; Templeton, J.G.C. *Retrial Queues*; Chapman & Hall: London, UK, 1997.
14. Artalejo, J.R.; Gomez-Corral, A. *Retrial Queueing Systems*; Springer: Berlin/Heidelberg, Germany, 2008.
15. Falin, G.: A survey of retrial queues. *Queueing Syst.* **1990**, *7*, 127–167. [CrossRef]
16. Yang, T.; Templeton, J. G.C. A survey on retrial queues. *Queueing Syst.* **1987**, *2*, 201–233. [CrossRef]
17. Gomez-Corral, A. A bibliographical guide to the analysis of retrial queues through matrix analytic techniques. *Ann. Oper. Res.* **2006**, *141*, 163–191. [CrossRef]
18. Artalejo, J.R. Accessible bibliography on retrial queues: Progress in 2000–2009. *Math. Comput. Model.* **2010**, *51*, 1071–1081. [CrossRef]
19. Kim, J.; Kim, B.: A survey of retrial queueing systems. *Ann. Oper. Res.* **2016**, *247*, 3–36. [CrossRef]
20. 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]
21. Breuer, L.; Dudin, A.; Klimenok, V. A retrial $BMAP/PH/N$ system. *Queueing Syst.* **2002**, *40*, 433–457. [CrossRef]
22. Somashekar, G.; Delasay, M.; Gandhi, A. Truncating multi-dimensional Markov chains with accuracy guarantee. In Proceedings of the 2022 30th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Nice, France, 18–20 October 2022; pp. 121–128.
23. Neuts, M.F.; Rao B.M. Numerical investigation of a multiserver retrial model. *Queueing Syst.* **1990**, *7*, 169–189. [CrossRef]
24. Bright, L.; Taylor, P.G. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stoch. Model.* **1995**, *11*, 497–525. [CrossRef]
25. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 2012.
26. Gantmakher, F.R. *The Matrix Theory*; Science: Moscow, Russia, 1967.
27. Dudina, O.; Kim C.; Dudin, S. Retrial queuing system with Markovian arrival flow and phase-type service time distribution. *Comput. Ind. Eng.* **2013**, *66*, 360–373. [CrossRef]
28. Dudin, S.; Dudina, O. Retrial multi-server queuing system with PHF service time distribution as a model of a channel with unreliable transmission of information. *Appl. Math. Model.* **2019**, *65*, 676–695. [CrossRef]
29. 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]
30. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Ellis Horwood: Cichester, UK, 1981.
31. Kim, C.S.; Mushko, V.V.; Dudin, A. Computation of the steady state distribution for multi-server retrial queues with phase type service process. *Ann. Oper. Res.* **2012**, *201*, 307–323. [CrossRef]
32. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer Nature: Cham, Switzerland, 2020.
33. Lucantoni, D. New results on the single server queue with a batch Markovian arrival process. *Commun. Stat. Stoch. Model.* **1991**, *7*, 1–46. [CrossRef]
34. 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.
35. O'Cinneide, C.A. Phase-type distributions: Open problems and a few properties. *Stoch. Model.* **1999**, *15*, 731–757. [CrossRef]
36. 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]