Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2026. Том 34. № 1. С. 118–124

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
PROBABILITY THEORY AND MATHEMATICAL STATISTICS

UDC 519.872 https://doi.org/10.67268/1812-5093-2026-34-1-118-124
EDN: PBYKZA

DETAILED PROOF OF ERGODICITY CONDITION FOR THE MULTI-SERVER RETRIAL QUEUEING SYSTEM WITH HETEROGENEOUS SERVERS AND PHASE TYPE DISTRIBUTION OF SERVICE

M. Liu, A. N. Dudin

Belarusian State University, Minsk, Belarus
e-mail: liumei19910101@126.com, dudin@bsu.by

Received: 05.02.2026 **Revised:** 20.03.2026 **Accepted:** 29.06.2026

Keywords: Markovian arrival process, retrials, heterogeneous servers, phase-type distribution, asymptotically quasi-Toeplitz Markov chains, ergodicity.

Abstract. Detailed proof of ergodicity for a multi-server retrial queueing system with heterogeneous servers, service times having a phase-type distribution with different irreducible representations and customer arrival defined by a Markovian arrival process is given. The proof consists of the use of the asymptotically quasi-Toeplitz Markov chains and Markov renewal processes theory.

ПОДРОБНОЕ ДОКАЗАТЕЛЬСТВО УСЛОВИЯ ЭРГОДИЧНОСТИ ДЛЯ МНОГОЛИНЕЙНОЙ СИСТЕМЫ МАССОВОГО ОБСЛУЖИВАНИЯ С НЕОДНОРОДНЫМИ ПРИБОРАМИ И РАСПРЕДЕЛЕНИЕМ ВРЕМЕНИ ОБСЛУЖИВАНИЯ ФАЗОВОГО ТИПА

Лю Мэй, А. Н. Дудин

Белорусский государственный университет, Минск, Беларусь
e-mail: liumei19910101@126.com, dudin@bsu.by

Поступила: 05.02.2026 **Исправлена:** 20.03.2026 **Принята:** 29.06.2026

Ключевые слова: марковский процесс поступления, повторные попытки, неоднородные серверы, распределение фазового типа, асимптотически квазитеплицевы цепи Маркова, эргодичность.

Аннотация. Представлено подробное доказательство условия эргодичности для многолинейной системы массового обслуживания с повторными попытками, неоднородными приборами, временем обслуживания, имеющим фазовое распределение с различными неприводимыми представлениями, и поступлением запросов, определяемым марковским процессом поступления. Доказательство состоит в использовании асимптотически квазитеплицевых цепей Маркова и теории марковских процессов обновления.

1. Introduction

Retrial queues are good mathematical models of a variety of important real-world systems. Therefore, their analysis attracts significant attention of many researchers. In [1], quite general multi-server retrial queueing system with a finite number of heterogeneous servers, service times having a phase-type distribution with different irreducible representations and customer arrival defined by a Markovian arrival process was analysed. The survey of the related research is provided there.

Essential step in analysis of any stochastic process is establishing conditions for ergodicity of the process. A random process is called ergodic if time averages of a single sample function (realization) are equal to the ensemble averages (the statistical average across all possible realizations). Equivalently, a random process is called ergodic if the statistical dependence between its values at two time points vanishes as the distance between these points increases to infinity. In application to queueing models, the ergodicity of a process describing the dynamics of the system implies the existence of so called stationary regime of the system operation or stationary (steady-state, invariant) distribution of the system states.

Stochastic processes describing the dynamics of multi-server retrial queues are state inhomogeneous what makes derivation of ergodicity condition for these queues very difficult. In [1], such a condition is presented for quite general retrial queue with Markovian arrival process, heterogeneous servers whose service times have so-called phase-type (*PH*) distribution, arbitrary dependence of the total retrial rate on the number of retrying customers, the priority for service beginning given to the servers with the minimal serial number (index) and impossibility to preempt service and change the server at which service is provided. Ergodicity condition has a simple analytically tractable form. However, its proof, which is based on the use of the results obtained for so-called asymptotically quasi-Toeplitz Markov chains, see [2], presented in [1] appears to be too brief. The goals of this paper are: (i) to present such a proof in more detailed form; and (ii) to propose the new possible way for simplification of ergodicity condition for a large class of multidimensional Markov chains under the presence of some specifics of the concrete queueing model.

2. The mathematical model

We present here a very brief description of the retrial queue under study. Detailed description can be found in [1].

– The system has N independent, generally speaking, non-identical servers. Servers are enumerated in some order. Admitted customer receives service in the server having the minimal number among all idle servers. After service beginning, transition of the customer to another server is not allowed.

– The service time of a customer by n-th server, $n = \overline{1,N}$, has *PH* distribution. It is governed by the continuous-time Markov chain (directing process) $\eta_t^{(n)}$, $t \geqslant 0$. This process has an absorbing state 0 and the set $\{1, \cdots, M^{(n)}\}$ of transient states. The initial state of the process $\eta_t^{(n)}$ at the epoch of starting the service is chosen among the transient states with the probabilities defined by the entries of the row-vector $\beta^{(n)} = (\beta_1^{(n)}, \ldots, \beta_{M^{(n)}}^{(n)})$. The transitions of the process $\eta_t^{(n)}$ inside the set of transient states do not lead to service completion and are defined by the entries of the irreducible matrix $S^{(n)}$ of size $M^{(n)}$. The diagonal entries of this matrix are negative. Their modules define the rates of the exit of the process $\eta_t^{(n)}$ from its transient states. The non-diagonal entries define the intensities of transitions inside the set of the transient states. The rates of transition to the absorbing state, which lead to service completion, are defined by the entries of the column vector $S_0^{(n)} = -S^{(n)}\boldsymbol{e}$.

The value μ_n defined by the formula $\mu_n^{-1} = \beta^{(n)}(-S^{(n)})^{-1}\boldsymbol{e}$, $n = \overline{1,N}$, is the mean service rate in the nth server. More information about the *PH* distribution and its properties can be found in [3].

– The customers arrive according to a *MAP* (Markovian Arrival Process). The underlying process of the *MAP*, ν_t, $t \geqslant 0$, is the irreducible continuous-time Markov chain having the state space $\{0, 1, \ldots, W\}$. The intensities of transitions of the process ν_t are defined as the entries of the square matrices D_0 and D_1 of size $\bar{W} = W + 1$. The matrix D_0 contains the intensities of transitions at which customers do not arrive. The matrix D_1 contains the intensities of transitions at which customer arrives to the system. The vector θ that is the unique solution to the system of equations $\theta(D_0 + D_1) = \mathbf{0}, \theta\boldsymbol{e} = 1$ defines the stationary distribution of the process ν_t. Here and thereafter $\boldsymbol{e}$ is a column vector of an appropriate size consisting of 1's and $\mathbf{0}$ is a row vector of an appropriate size consisting of zeroes.

The average arrival rate λ of the *MAP* is defined as $\lambda = \theta D_1\boldsymbol{e}$.

More information about the *MAP* arrival flow, its properties, formulas for computation of the coefficients of correlation and variation as well as higher moments of distribution of inter-arrival times, important particular cases, possible generalizations and usefulness for modelling the correlated bursty flows in modern service systems, telecommunication networks in particular, can be found in the book [3].

– If the arriving customer meets idle servers, it immediately starts service in the server with the minimal number among available servers. If all servers are busy, then the arriving customer goes to the virtual place called orbit. Capacity of the orbit is unlimited. These customers are said to be repeated customers. These customers try their luck later until they will be served. We assume that the total flow of retrials from the orbit is such that the probability of generating the retrial attempt in the small interval

$(t, t+\Delta t)$ is equal to $\alpha_i \Delta t + o(\Delta t)$ when the orbit size (the number of customers on the orbit) is equal to i, $i > 0$, $\alpha_0 = 0$. We assume the infinitely increasing retrial rate: $\lim\limits_{i\to\infty} \alpha_i = \infty$.

3. The random process defining the dynamics of the system

Description of the dynamics of any multi-server retrial queue with the *MAP* arrivals and *PH* distribution of service times includes information about the number, say i_t, of retrying customers at the moment t, $t > 0$, where $i_t \geqslant 0$. Account of *MAP* arrivals implies permanent monitoring the state ν_t of the underlying process of the arrivals, $\nu_t = \overline{0,W}$. Additionally, account of the Markov processes with a finite state describing simultaneous service of admitted customers by the servers is mandatory. If servers are homogeneous and service times are exponentially distributed, the service process is completely defined by the number, say n_t, of busy servers. If servers are homogeneous and service times have more general, *PH*-type distribution, besides the number n_t of busy servers, it is necessary to specify the current phase of service at each busy server or the number of customers receiving service at any phase. In the considered in this paper system, the servers are heterogeneous. Therefore, for each server it is necessary to specify whether the server is busy or idle. In the former case, the current phase of service has to be specified.

Having in mind the presented considerations, we describe the dynamics of the system under study by the following multidimensional Markov process:

$$\zeta_t = \{i_t, \eta_t^{(1)}, \ldots, \eta_t^{(N)}, \nu_t\}, \ t \geqslant 0,$$

where $\eta_t^{(n)}$ is the state of the underlying process of the service in the nth server, $n = \overline{1,N}$. This state belongs to the set $\{1, \ldots, M^{(n)}\}$ if this server is busy and is assumed to be 0 if the server is idle.

Let us enumerate the states of the Markov chain ζ_t in the lexicographic order and call the set of the states of the chain having the value i of the component i_t, as level i, $i \geqslant 0$. Let us combine the transition rates from the level i to level j into the square matrices $\mathbf{Q}_{i,j}$ of size $\hat{M}\bar{W}$, where $\hat{M} = \prod\limits_{n=1}^{N} (M^{(n)} + 1)$. Here, $\max\{i-1, 0\} \leqslant j \leqslant i+1$, $i \geqslant 0$.

The following formulas for the matrices $\mathbf{Q}_{i,j}$ are derived in [1]:

$$\mathbf{Q}_{i,i+1} = J \otimes D_1, \ \mathbf{Q}_{i,i-1} = \alpha_i \tilde{I}_\beta \otimes I_{\bar{W}}, \ \mathbf{Q}_{i,i} = G \oplus D_0 - \alpha_i \bar{I} \otimes I_{\bar{W}} + \tilde{I}_\beta \otimes D_1.$$

Here: matrix J is defined as

$$J = \bigotimes_{n=1}^{N} J_n, \ J_n = \begin{pmatrix} O_{1\times 1} & O_{1\times M^{(n)}} \\ O_{M^{(n)}\times 1} & I_{M^{(n)}} \end{pmatrix}, n = \overline{1,N};$$

matrix G is defined as

$$G = \sum_{n=1}^{N} \left(I_{\prod\limits_{l=1}^{n-1}(M^{(l)}+1)} \otimes G_n \otimes I_{\prod\limits_{l=n+1}^{N}(M^{(l)}+1)} \right),$$

where the matrix G_n, $n = \overline{1,N}$, is defined as

$$G_n = \begin{pmatrix} O_{1\times 1} & O_{1\times M^{(n)}} \\ S_0^{(n)} & S^{(n)} \end{pmatrix};$$

matrix Γ_n is defined as

$$\Gamma_n = \bigotimes_{l=1}^{n-1} J_l \otimes G_n \otimes \bigotimes_{l=n+1}^{N} J_l, \ n = \overline{1,N};$$

matrix $\tilde{I}_\beta$ is defined as

$$\tilde{I}_\beta = \sum_{n=1}^{N} \left(\bigotimes_{k=1}^{n-1} J_k \otimes \begin{pmatrix} O_{1\times 1} & \boldsymbol{\beta}^{(n)} \\ O_{M^{(n)}\times 1} & O_{M^{(n)}\times M^{(n)}} \end{pmatrix} \otimes I_{\prod\limits_{k=n+1}^{N}(M^{(k)}+1)} \right);$$

$$\bar{I} = (I_{\hat{M}} - J);$$

$\otimes$ and $\oplus$ are the symbols of the Kronecker product and sum of matrices, see, e. g., [4].

Derivation of ergodicity conditions for the Markov chain ζ_t is based on the use of the corresponding results for the asymptotically quasi-Toeplitz Markov chains given in [2]. According to [2], it is necessary to consider the discrete-time multidimensional Markov chain describing transitions of the states of the continuous-time Markov chain ζ_t at all moments of the jumps of the chain ζ_t.

For this embedded Markov chain, it is proven in [1] that the following limits exist:

$$Y_0 = \lim_{i\to\infty} R_i^{-1}\mathbf{Q}_{i,i-1},\; Y_2 = \lim_{i\to\infty} R_i^{-1}\mathbf{Q}_{i,i+1},\; Y_1 = \lim_{i\to\infty} R_i^{-1}\mathbf{Q}_{i,i} + I,$$

where $R_i = -I \circ \mathbf{Q}_{i,i} = \alpha_i \bar{I} \otimes I_{\bar{W}} + C,\; i \geqslant 0$, where $C = -I \circ (G \oplus D_0)$ and $\circ$ denotes Hadamard product of matrices, see [5].

By direct calculations, it was shown in [1] that the matrices $Y_k,\; k = 0,1,2$, are defined by formulas

$$Y_0 = \tilde{I}_\beta \otimes I_{\bar{W}},\; Y_2 = C^{-1}(J \otimes D_1),\; Y_1 = C^{-1}(\sum_{n=1}^{N} \Gamma_n \oplus D_0) + J \otimes I_{\bar{W}}.$$

4. Proof of ergodicity condition

The following assertion is given in [1].

Theorem 4.1. *The following statements hold good:*

(i) *The Markov chain ζ_t is ergodic if the inequality*

$$\lambda < \sum_{k=1}^{N} \mu_k \tag{1}$$

is fulfilled.

(ii) *The Markov chain ζ_t is non-ergodic if*

$$\lambda > \sum_{k=1}^{N} \mu_k.$$

Proof. According to [2], the Markov chain ζ_t is ergodic if the inequality

$$\mathbf{y}Y_2\boldsymbol{e} < \mathbf{y}Y_0\boldsymbol{e}, \tag{2}$$

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

$$\mathbf{y}(Y_0 + Y_1 + Y_2) = \mathbf{y},\;\; \mathbf{y}\boldsymbol{e} = 1 \tag{3}$$

is fulfilled.

The Markov chain ζ_t is non-ergodic if inequality (2) has an opposite sign.

Condition (2) is easily verified on computer for any asymptotically quasi-Toeplitz Markov chain, once the limiting matrices $Y_k,\; k = 0,1,2$, are computed and a finite system (3) is solved.

However, sometimes, when the generator $\mathbf{Q}$ of the chain has some specifics, it is possible to reduce inequality (2) to the simple and transparent scalar form like (1) in the model under study in this paper. To implement such a reduction, it is necessary to solve system (3) not numerically, but explicitly.

If the servers of the multi-server retrial queue are identical, the levels i can be evidently partitioned into sublevels $((i,0),(i,1),\dots,(i,N))$ where sublevel (i,n) contains the states of the Markov chain when i customers stay in orbit and n servers are busy. Correspondingly, vector $\mathbf{y}$, which is solution of system (3), is partitioned as $\mathbf{y} = (\mathbf{y}_0,\mathbf{y}_1,\dots,\mathbf{y}_N)$. Using such a presentation, often system (3) can be solved explicitly. Matrix of this system is reducible. This causes that the vectors $\mathbf{y}_n,\; n = \overline{0,N-2}$, are equal to zero. The vectors $\mathbf{y}_n,\; n = N-1,N$ are then found as solution of a small subsystem of system (3).

However, in analysis of the queueing system under study, such a natural decomposition of the vector $\mathbf{y}$ is not possible due to the complicated structure of the matrix $(Y_0 + Y_1 + Y_2)$ of this system which excludes the possibility of the partition $\mathbf{y} = (\mathbf{y}_0,\mathbf{y}_1,\dots,\mathbf{y}_N)$. Therefore, we cannot solve the system (3) by means of the algebraic manipulations as in the case of homogeneous servers. Instead, we use the probabilistic considerations.

Let us denote by $\mathbf{u}_n,\ n=\overline{0,N}$, the vectors obtained from the vector $\mathbf{y}$ by setting equal to zero all the components that correspond to the states of the Markov chain ζ_t such that the number of busy servers is not equal to n. In other words, the vector $\mathbf{u}_n$ includes only the corresponding entries of the vector $\mathbf{y}$ for which the number of the components $\eta_t^{(l)},\ l=\overline{1,N}$, such that $\eta_t^{(l)}\neq 0$ is equal to n.

It is obvious, that

$$\mathbf{y}=\sum_{n=0}^{N}\mathbf{u}_n.$$

According to definition in [2], the vector $\mathbf{y}$ defines the stationary distribution of the discrete-time Markov chain that is embedded at the moments immediately after the jumps of the continuous-time Markov chain ζ_t when the value of the component i_t approaches infinity. Because when the component i_t approaches infinity, i.e., the number of customers in the orbit becomes infinite, the new service begins immediately after service completion in any server. Therefore, with the positive probability, the number of busy servers at the embedded moment can be equal to N (when no transitions of the underlying process of arrivals or underlying processes of service, which imply the service completion, occur at the jump moment) or equal to $N-1$ when the transition of the underlying processes of service, which leads to service completion in one of the busy servers, occurs. In the latter case, new service immediately starts in the just released server.

As follows from the presented consideration, the vectors $\mathbf{u}_n$ are zero vectors for $n=\overline{0,N-2}$, and, therefore, the vector $\mathbf{y}$, which is the solution of system (3), is defined by

$$\mathbf{y}=\mathbf{u}_N+\mathbf{u}_{N-1}. \tag{4}$$

Let us calculate the vectors $\mathbf{u}_N$ and $\mathbf{u}_{N-1}$. It is known, see, e. g., [3;6], that if the service in the nth server is providing permanently, i. e., new service begins immediately after service completion, then the steady-state distribution of the underlying process of service in this server is defined by the vector

$$\psi_n=\mu_n\beta^{(n)}(-S^{(n)})^{-1},\ n=\overline{1,N},$$

and the joint steady-state distribution of the underlying processes of service in the system with permanently busy N servers is defined by the vector

$$\psi=\bigotimes_{n=1}^{N}(0,\psi_n)$$

which is the solution of the system

$$\psi\sum_{n=1}^{N}\left[\bigotimes_{k=1}^{n-1}J_k\otimes\begin{pmatrix}O_{1\times 1} & O_{1\times M^{(n)}}\\ O_{M^{(n)}\times 1} & S^{(n)}+\mathbf{S}_0^{(n)}\beta^{(n)}\end{pmatrix}\otimes\bigotimes_{k=n+1}^{N}J_k\right]=\mathbf{0},\ \psi\boldsymbol{e}=1.$$

Therefore, the vector of the stationary probabilities of these underlying processes and, independent of them, underlying process of arrivals at an arbitrary moment is equal to $\psi\otimes\theta$ where, as it is defined above, θ is the vector of the steady state probabilities of the underlying process of arrivals.

As follows from the theory of Markov renewal processes, see, e. g., [3; 7], the vectors of the stationary probabilities of all underlying processes at an arbitrary moment are expressed via the vectors of the stationary probabilities of these processes at the jump moments as follows:

$$\psi\otimes\theta=\tau^{-1}\mathbf{u}_N\int_0^\infty e^{I\circ Gx}\otimes e^{I\circ D_0x}dx=\tau^{-1}\mathbf{u}_N\int_0^\infty e^{I\circ(G\oplus D_0)x}dx \tag{5}$$

where τ is the average length of the interval between the jumps, the matrix $e^{I\circ Gx}$ defines the probability that the underlying processes of the service make no transition during time x and the matrix $e^{I\circ D_0x}$ defines the probability that the underlying process of the arrivals makes no transition during time x.

By calculating the integral, what is possible because the matrix $G\oplus D_0$ is the irreducible subgenerator, relation (5) is rewritten as

$$\psi\otimes\theta=\tau^{-1}\mathbf{u}_NC^{-1},$$

what is equivalent to

$$\mathbf{u}_N = \tau(\boldsymbol{\psi}\otimes\boldsymbol{\theta})C. \tag{6}$$

Formula for computation of the vector $\mathbf{u}_{N-1}$ is easily derived using the following consideration. Transition of the embedded Markov chain to the state where $N-1$ servers are busy is possible from the state when N servers are busy after some time, say, x, $x\geqslant 0$, during which no transitions occur, via service completion in one of busy servers during the small interval $(x, x+dx)$. The rates of the service completion in one of nth busy servers, conditional that the system is overloaded, i. e., all other servers are busy as well and new service immediately starts in the released server, are given by the matrix $\tilde{\Gamma}_n = \bigotimes_{l=1}^{n-1} J_l \otimes \tilde{G}_n \otimes \bigotimes_{l=n+1}^{N} J_l$, $n = \overline{1,N}$, where the matrix $\tilde{G}_n$ is defined as

$$\tilde{G}_n = \begin{pmatrix} O_{1\times 1} & O_{1\times M^{(n)}} \\ \mathbf{0}^T & S_0^{(n)}\boldsymbol{\beta}^{(n)} \end{pmatrix},\ n=\overline{1,N}.$$

Here $\mathbf{0}^T$ is the transposed row vector $\mathbf{0}$.

As the result of these considerations, we obtain formula

$$\mathbf{u}_{N-1} = \mathbf{u}_N \int_0^\infty e^{I\circ(G\oplus D_0)x}\Big(\sum_{n=1}^{N}\tilde{\Gamma}_n\otimes I_{\bar{W}}\Big)dx$$

or, finally,

$$\mathbf{u}_{N-1} = \mathbf{u}_N C^{-1}\Big(\sum_{n=1}^{N}\tilde{\Gamma}_n\otimes I_{\bar{W}}\Big) = \tau(\boldsymbol{\psi}\otimes\boldsymbol{\theta})\Big(\sum_{n=1}^{N}\tilde{\Gamma}_n\otimes I_{\bar{W}}\Big). \tag{7}$$

Formulas (4), (6), and (7) completely define the vector $\mathbf{y}$ that is the solution of system (3). Note that the positive constant τ can be easily found from the normalization condition $(\mathbf{u}_{N-1}+\mathbf{u}_N)\mathbf{e} = 1$, but we do not need to know this constant in further derivation.

Now, we have to derive inequality (1) from inequality (2). It follows from (4) that inequality (2) can be rewritten as

$$\mathbf{u}_N Y_2\mathbf{e} + \mathbf{u}_{N-1}Y_2\mathbf{e} < \mathbf{u}_N Y_0\mathbf{e} + \mathbf{u}_{N-1}Y_0\mathbf{e}.$$

This inequality is reduced to a simpler form

$$\mathbf{u}_N Y_2\mathbf{e} < \mathbf{u}_{N-1}Y_0\mathbf{e} \tag{8}$$

because the matrix Y_2 has non-zero rows only for the states of the embedded Markov chain corresponding to all busy servers while the vector $\mathbf{u}_{N-1}$ has non-zero components corresponding to $N-1$ busy servers; the matrix Y_0 has non-zero rows only for the states of the embedded Markov chain when at least one server is idle while the vector $\mathbf{u}_N$ has non-zero components only for the states with N busy servers.

The left hand side of inequality (8) is transformed as follows:

$$\mathbf{u}_N Y_2\mathbf{e} = \tau(\boldsymbol{\psi}\otimes\boldsymbol{\theta})(J\otimes D_1)\mathbf{e} = \tau(\boldsymbol{\psi}J\otimes\boldsymbol{\theta}D_1)\mathbf{e} = \tau(\boldsymbol{\psi}J)\mathbf{e}\otimes(\boldsymbol{\theta}D_1)\mathbf{e} = \tau\lambda. \tag{9}$$

The so-called mixed product rule for the Kronecker products of matrices ($(AB)\otimes(CD) = (A\otimes C)(B\otimes D)$ for matrices A,B,C,D of matching sizes) was used in transformations (9) along with the definition of the mean arrival rate λ.

The right hand side of inequality (8) is transformed as follows:

$$\mathbf{u}_{N-1}Y_0\mathbf{e} = \tau(\boldsymbol{\psi}\otimes\boldsymbol{\theta})\Big(\sum_{n=1}^{N}\tilde{\Gamma}_n\otimes I_{\bar{W}}\Big)\mathbf{e} = \tau\boldsymbol{\psi}\Big(\sum_{n=1}^{N}\tilde{\Gamma}_n\Big)\mathbf{e} =$$

$$= \tau\bigotimes_{n=1}^{N}(0,\boldsymbol{\psi}_n)\Big(\sum_{n=1}^{N}\bigotimes_{l=1}^{n-1}J_l\otimes\tilde{G}_n\otimes\bigotimes_{l=n+1}^{N}J_l\Big)\mathbf{e} = \tau\sum_{n=1}^{N}\mu_n. \tag{10}$$

Inequality (1) evidently follows from formulas (8), (9), and (10). Theorem is proven. □

References

1. Liu M., Dudin A. N. Steady-state analysis of the multi-server retrial queueing system with heterogeneous servers and phase type distribution of service times. *Proceedings of the Institute of Mathematics of the NAS of Belarus*, 2025, vol. 33, no. 1, pp. 111–120.

2. Klimenok V. I., Dudin A. N. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*, 2006, vol. 54, pp. 245–259.

3. Dudin A. N., Klimenok V. I., Vishnevsky V. M. *The Theory of Queuing Systems with Correlated Flows*. Springer Nature, 2020.

4. Graham A. *Kronecker Products and Matrix Calculus with Applications*. Courier Dover Publications, Mineola, 2018.

5. Horn R. A., Johnson C. R. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 2012.

6. Neuts M. *Matrix-Geometric Solutions in Stochastic Models*. North Chelmsford, Courier Corporation, 1994.

7. Cinlar E. *Introduction to Stochastic Processes*. Courier Corporation, 2013.