Национальная академия наук Беларуси
Труды Института математики НАН Беларуси. 2025. Том 33. № 1. С. 111–120

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА PROBABILITY THEORY AND MATHEMATICAL STATISTICS

UDC 511.42

STEADY-STATE ANALYSIS OF THE MULTI-SERVER RETRIAL QUEUEING SYSTEM WITH HETEROGENEOUS SERVERS AND PHASE TYPE DISTRIBUTION OF SERVICE TIMES

M. Liu, A. N. Dudin

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

Received: 21.04.2025 **Revised:** 23.05.2025 **Accepted:** 23.05.2025

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

Abstract. A multi-server retrial queueing system with heterogeneous servers is analysed. The service times have a phase-type distribution with different irreducible representations. Customer arrival to the system is defined by a Markovian arrival process. When all servers are busy at an arrival moment, the customer moves to the virtual place called orbit to retry to reach the servers in exponentially distributed periods of time. The total retrial rate from the orbit infinitely increases when the number of customers residing in orbit grows. Upon arrival or retrial from the orbit, a customer occupies the server having the minimal number among all idle servers, if any. The dynamics of the system states is described by a multidimensional Markov chain having the special block structure of the infinitesimal generator. The explicit expression for this is presented. Ergodicity condition is derived. The expressions for computation of the key performance characteristics of the system are given. Numerical results, which highlight dependencies of these measures on the mean arrival rate for the system and its particular cases, when the arrivals are described by the stationary Poisson process or (and) service times follow the exponential distribution, are presented.

СТАЦИОНАРНЫЙ АНАЛИЗ 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: 21.04.2025 **Исправлена:** 23.05.2025 **Принята:** 23.05.2025

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

Аннотация. Анализируется многолинейная система обслуживания с повторными вызовами и неоднородными серверами. Длительности обслуживания имеют фазовое распределение с различными неприводимыми представлениями. Поступление запросов в систему определяется марковским процессом поступления. Когда все серверы заняты в момент поступления, запрос помещается в виртуальное место, называемое орбитой, чтобы повторить попытку достичь серверов через экспоненциально распределенные периоды времени. Общая скорость повторных вызовов с орбиты бесконечно увеличивается с ростом числа запросов, находящихся на орбите. При поступлении или повторных вызовах с орбиты запрос занимает сервер с минимальным номером среди всех свободных серверов, если таковые имеются. Динамика состояний системы описывается многомерной цепью Маркова, имеющей специальную блочную структуру инфинитезимального генератора. Представлено явное выражение для генератора. Выведено условие эргодичности. Приведены выражения для вычисления ключевых характеристик производительности системы. Представлены численные результаты, иллюстрирующие зависимости характеристик производительности системы от средней скорости поступления заявок для системы и ее частных случаев, когда поступления описываются стационарным пуассоновским процессом или (и) времена обслуживания подчиняются экспоненциальному распределению.

1. Introduction

Retrial queues fit well for the mathematical description of various real-world systems, telecommunication networks, including the mobile cellular networks, and contact centers in particular. Analysis of such queues is much more involved than the study of the queues with customer loss or buffers for waiting in case of the lack of available servers having the same types of arrival and service processes. This is explained by to the state inhomogeneous behavior of the stochastic process describing the dynamics of the system. This is the reason why the retrial queues are investigated in a far less extent.

The fundamental results obtained for the multi-server retrial queues of the $M/M/N$ type (this means that the arrivals are described by the stationary Poisson process and service times follow the exponential distribution) are presented in the well-known [1]. However, due to the significant change of the character of the flows in communication networks and customers service time during the last decades, essentially more adequate model of arrivals in the modern real-world systems is the *MAP* (Markovian Arrival Process), see, e. g., [2–5]. This process well describes the modern correlated bursty flows. Essentially more general distribution of service time than the exponential one is Phase-Type (*PH*) distribution, see, e. g., [5; 6] which allows to fit not only the mean service time, but also higher moments, including the variance. Taking these circumstances into account, the $BMAP/PH/N$ type retrial queue was studied in [7]. But only the aspects relating to the ergodicity condition of the system are considered there. More comprehensive analysis of the $BMAP/PH/N$ type retrial queue was given [8] where, besides to the proof of the sufficient condition for the ergodicity in cases of the classical retrial policy and the constant retrial rate, the effective algorithms for the computation of the stationary distribution of the system states and the main performance measures were presented.

Essential assumption imposed in [1] and [8] is that the servers are identical, while they can be heterogeneous in some real-world system. In this paper, we significantly weaken this restrictive assumption. In the papers [9] and [10], this assumption was already weakened and the servers are not identical. In that papers, it was supposed that service times have the exponential distributions with different parameters. In the present paper, we suppose that service times have the Phase-Type distributions with different parameters.

Generalization from the case of the exponential distribution to the case of the Phase-Type distribution has the practical importance because the former one allows to fit only the average value of the real service time while the latter one allows to fit simultaneously many initial moments of the distribution of the real service time, and the variance of this time in particular. From the theoretical point of view, this generalization leads to the necessity of construction and analysis of the multi-dimensional continuous-time Markov chain with more involved structure of the blocks on the infinitesimal generator. Here this analysis is successfully implemented.

The outline of the presentation is as follows. The mathematical model under study is described in Section 2. The random process describing the dynamics of the considered system is introduced in Section 3 as the multi-dimensional continuous-time Markov chain and the explicit expression for the generator of the chain is presented there. The sufficient condition for the ergodicity of this Markov chain is derived in Section 4. Formulas for computation of the values of the key performance measures of the system are given in Section 5. The illustrative numerical results are presented in Section 6. Section 7 contains some concluding remarks.

2. The mathematical model

We consider an N-server queueing system. The primary customers arrive to the system according to a *MAP* (Markovian Arrival Process). We denote the directing process of the *MAP* by ν_t, $t \geqslant 0$. The state space of the irreducible continuous-time Markov chain ν_t is $\{0, 1, \ldots, W\}$. 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 matrix $D(1) = D_0 + D_1$ is an infinitesimal generator of the process ν_t. The vector θ that is the unique solution to the system of equations $\theta D(1) = \mathbf{0}$, $\theta \boldsymbol{e} = 1$ 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 (fundamental) arrival rate λ of the *MAP* is defined as $\lambda = \theta D_1 \boldsymbol{e}$. The coefficient c_{var} of variation of intervals between customer arrivals is defined by $c_{\text{var}} = 2\lambda\theta(-D_0)^{-1}\boldsymbol{e} - 1$. The coefficient of correlation c_{var} of successive intervals between arrivals is computed as $c_{\text{cor}} = (\lambda\theta(-D_0)^{-1}D_1(-D_0)^{-1}\boldsymbol{e} - -1)/c_{\text{var}}^2$.

The servers are independent of each other. The service time of a customer by n-th server, $n = \overline{1,N}$, is governed by the continuous-time Markov chain (directing process) $\eta_t^{(n)}$. This process has an absorbing state 0 and the set $\left\{1,\cdots,M^{(n)}\right\}$ 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 $\boldsymbol{\beta}^{(n)} = (\beta_1^{(n)},...,\beta_{M^{(n)}}^{(n)})$. 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 $S_0^{(n)} = -S^{(n)}\mathbf{e}$.

The mth initial moment $b_m^{(n)}$ of the distribution of the service time in the nth server is computed as

$$b_m^{(n)} = m!\boldsymbol{\beta}^{(n)}((-S)^{(n)})^{-m}\boldsymbol{e},\ m \geqslant 1.$$

The value μ_n defined by the formula $\mu_n^{-1} = \boldsymbol{\beta}^{(n)}(-S^{(n)})^{-1}\boldsymbol{e}$, $n = \overline{1,N}$, is the mean service rate in the nth server. The value $\frac{b_2^{(n)} - (b_1^{(n)})^2}{(b_1^{(n)})^2}$ is the squared coefficient of variation of the service time in the nth server.

If the arriving customer meets all servers being idle, the customer enters the first server to receive the service. If the first server is busy, then the customer enters the idle server with the minimum number. If all servers are busy, then the 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, $i > 0$, $\alpha_i = 0$. We do not fix the explicit dependence of the intensities α_i on i. We assume the infinitely increasing retrial rate: $\lim\limits_{i\to\infty} \alpha_i = \infty$. This holds true, in particular, for classic retrial strategy where $\alpha_i = i\alpha$ and the linear strategy $\alpha_i = i\alpha + \gamma$.

Our goal is to obtain the sufficient condition for existence of stationary state distribution of the system, this distribution itself, and the expressions for the key performance measures of the system via this stationary distribution.

3. The random process defining the dynamics of the system

Let, at the moment t, $t > 0$,

i_t be the number of customers on the orbit, $i_t \geqslant 0$;

$\eta_t^{(n)}$ be 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;

ν_t be the state of the directing process of the *MAP*, $\nu_t = \overline{0,W}$.

Let $\mathcal{R}$ be the state space of the process $\{\eta_t^{(1)},\ldots,\eta_t^{(N)}\}$ defining the phases of service in all servers of the system:

$$\mathcal{R} = \{(r^{(1)},\ldots,r^{(N)}) :\ r^{(n)} = \overline{0,M^{(n)}},\ n = \overline{1,N}\}.$$

Consider the continuous time multi-dimensional process

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

It is easy to see that this process is an irreducible continuous-time Markov chain.

Let us define the stationary probabilities of this Markov chain as the limits

$$\pi(i,r^{(1)},\dots,r^{(N)},\nu)=\lim_{t\to\infty}P\{i_t=i,(\eta_t^{(1)},\dots,\eta_t^{(N)})=(r^{(1)},\dots,r^{(N)})\in\mathcal{R},\nu_t=\nu\},$$

$$i\geqslant 0,\ \nu=\overline{0,W}.$$

Sufficient condition for existence of these limits will be presented in Theorem 4.1 below.

Let us enumerate the states of the chain ζ_t, $t\geqslant 0$, in the lexicographic order and form the row-vector

$$\boldsymbol{\pi}(i,r^{(1)},\dots,r^{(N)})=(\pi(i,r^{(1)},\dots,r^{(N)},0),\dots,\pi(i,r^{(1)},\dots,r^{(N)},W))$$

of the stationary probabilities $\pi(i,r^{(1)},\dots,r^{(N)},\nu)$, and the row-vectors $\boldsymbol{\pi}_i$, consisting of the vectors $\boldsymbol{\pi}(i,r^{(1)},\dots,r^{(N)})$, $i\geqslant 0$.

Note that the size of the vectors $\boldsymbol{\pi}_i$ is equal to $K=(W+1)\hat{M}$ where $\hat{M}=\prod_{n=1}^{N}(M^{(n)}+1)$.

Define also the infinite-dimensional probability vector $\boldsymbol{\pi}=(\boldsymbol{\pi}_0,\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\dots)$.

For the use in the sequel, introduce the following notation:

I is an identity matrix of appropriate dimension (when needed, the dimension is identified with a subscript);

$O_{n\times n'}$ denotes zero matrices with n rows and n' columns; $\otimes$ and $\oplus$ are the symbols of the Kronecker product and sum of matrices, see, e. g., [11];

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

$\bigotimes_{l=r}^{m}J_l=J_r\otimes J_{r+1}\otimes\dots\otimes J_m,\ r\leqslant m,\ m=\overline{1,N};\ J=\bigotimes_{l=1}^{N}J_l$;

$\mathbf{f}^{(n)}$ is column vector of size $(M^{(n)}+1)$ having the first entry equal to 1 and other entries equal to 0;

$\delta_{i,j}$ is Kronecker delta. It is equal to 1 if $i=j$ and 0, otherwise;

Λ is the diagonal matrix with diagonal entries defined by the diagonal entries of the matrix D_0;

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

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

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

H_n is the diagonal matrix with the diagonal entries coinciding with the corresponding diagonal entries of the matrix G_n;

$H=\sum_{n=1}^{N}\left(\bigotimes_{l=1}^{n-1}J_l\otimes H_n\otimes\bigotimes_{l=n+1}^{N}J_l\right)$;

$C=-(H\oplus\Lambda)^{-1}$. The matrix $H\oplus\Lambda$ is nonsingular as the irreducible sub-generator;

$B_n=\begin{pmatrix}O_{1\times 1} & \boldsymbol{\beta}^{(n)}\\ O_{M^{(n)}\times 1} & O_{M^{(n)}\times M^{(n)}}\end{pmatrix},n=\overline{1,N}$;

$\tilde{I}_{\beta}=\sum_{k=1}^{N}\left(\bigotimes_{l=1}^{k-1}J_l\otimes B_k\otimes I_{\prod_{l=k+1}^{N}(M^{(l)}+1)}\right)$;

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

the product of numbers $\prod_{l=a}^{b}c_l$ or matrices $\prod_{l=a}^{b}C_l$ is supposed to be equal to 1 or I, correspondingly, if $b<a$. The same relates to the Kronecker products.

The following statements hold true:

Lemma 3.1. *If the vector $\boldsymbol{\pi}$ of stationary probabilities exists, then it satisfies the system of equilibrium equations*

$$\boldsymbol{\pi}\mathbf{Q}=\mathbf{0},\boldsymbol{\pi}\mathbf{e}=1$$

where $\mathbf{0}$ *is the infinite-size row-vector consisting of zeroes and the matrix* $\mathbf{Q}$, *which is the infinitesimal generator of the chain* $\zeta_t, t \geqslant 0$, *has the following structure:*

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{00} & \mathbf{Q}_{01} & O & O & \cdots \\ \mathbf{Q}_{10} & \mathbf{Q}_{11} & \mathbf{Q}_{12} & O & \cdots \\ O & \mathbf{Q}_{21} & \mathbf{Q}_{22} & \mathbf{Q}_{23} & \cdots \\ O & O & \mathbf{Q}_{32} & \mathbf{Q}_{33} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where the blocks $\mathbf{Q}_{i,j}, i, j \geqslant 0, j = \{\max\{0, i-1\}, i, i+1\}$, *of the matrix* $\mathbf{Q}$ *have size K and are defined as follows:*

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

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

Proof. The presented form of the blocks $\mathbf{Q}_{i,j}$ is easy explained if the intuitive meaning of some denotations is taken into account. In particular:

the matrix J_n is used to distinguish the residing of the process $\eta_t^{(n)}$ in the absorbing and transient states during which service is not provided and provided, correspondingly;

the matrix J highlights the states of the vector process $\boldsymbol{\eta}_t = \{\eta_t^{(1)}, \eta_t^{(2)}, \ldots, \eta_t^{(N)}\}$ at which all servers are busy;

the matrix $\bar{I}$ highlights the states of the vector process $\boldsymbol{\eta}_t$ at which not all servers are busy;

the matrix G_n describes transition rates of the process $\eta_t^{(n)}$ in its state space;

the matrix G contains the rates of all possible transitions of the process $\boldsymbol{\eta}_t$;

the matrix B_n is used to instal the initial state of the underlying process $\eta_t^{(n)}$ at the service beginning moment in the nth server;

the matrix $\tilde{I}_\beta$ describes transition rates of the process $\boldsymbol{\eta}_t$ at the moment of service beginning at the available server having the minimal number. Here, the matrix B_k defines the installment of service namely in the kth server, $k = \overline{1,N}$. This matrix is preceded by the Kronecker product $\bigotimes_{l=1}^{k-1} J_l$ that guarantees that the previous $k-1$ servers are busy and service cannot start in these servers. Additionally this matrix is multiplied from the right in Kronecker manner by the Kronecker product $I_{\prod_{l=k+1}^{N}(M^{(l)}+1)}$ that shows that no transitions occur in underlying processes of customers service in the servers having the numbers $k+1, k+2, \ldots, N$.

The increase in the number of customers in orbit occurs at the moment of a customer arrival to the system (with the rates defined by the entries of the matrix D_1) when all servers are busy. Therefore, we evidently obtain that $\mathbf{Q}_{i,i+1} = J \otimes D_1$.

The decrease in the number of customers in orbit occurs at the moment when one of the customers staying in orbit retries to enter the service (with the rate α_i if i customers stay in the orbit) when there are available servers and the server with the minimal number is occupied. The transition rates of the process $\boldsymbol{\eta}_t$ in this scenario are defined by the matrix $\tilde{I}_\beta$. Any transition in the underlying process of arrivals is not possible. Therefore, we evidently obtain that $\mathbf{Q}_{i,i-1} = \alpha_i \tilde{I}_\beta \otimes I_{\bar{W}}$.

The diagonal entries of the diagonal blocks $\mathbf{Q}_{i,i}$ are negative. Their modules define the exit rate of the Markov chain ζ_t, $t \geqslant 0$, from the corresponding states. The non-diagonal entries of the diagonal blocks $\mathbf{Q}_{i,i}$ are non-negative. They define the rates of the Markov chain ζ_t transitions that maintain the value of the number i of customers in the orbit. There exist four scenarios of such exits or transitions.

One scenario corresponds to the exit or transition of the underlying process of arrivals the rates of which are defined by the matrix D_0. No transition of the process $\boldsymbol{\eta}_t$ are allowed in this scenario. This scenario explains the first summand $I_{\hat{M}} \otimes D_0$ in the expression for $\mathbf{Q}_{i,i}$.

The second scenario of the exit from the current state due to successful retrial of a customer from the orbit. This scenario explains the second summand $-\alpha_i \bar{I} \otimes I_{\overline{W}}$ in the expression for $\mathbf{Q}_{i,i}$.

The third scenario corresponds to the exit or transition of the underlying process η_t of service, the rates of which are defined by the matrix G. No transition of the process ν_t are allowed in this scenario. This scenario explains the third summand in the form $G \otimes I_{\bar{W}}$.

The last scenario of transition of the Markov chain ζ_t corresponds to the new customer arrival, the rates of which are defined by the matrix D_1 and an immediate admission of this customer for service. Transitions probabilities of the process η_t at this arrival moment are defined by the matrix $\tilde{I}_{\beta}$. The proof of the formula for the block $\mathbf{Q}_{i,i}$ and of Lemma 3.1 is finished. □

Lemma 3.2. *Markov chain ξ_t belongs to the class of asymptotically quasi-Toeplitz Markov chains, see* [12].

Proof. According to the definition of the asymptotically quasi-Toeplitz Markov chains given in [12], we have to prove the existence of the limits

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

where R_i is a diagonal matrix with diagonal entries defined as the moduli of the corresponding diagonal entries of the matrix $\mathbf{Q}_{i,i}$, $i \geqslant 0$. It can be easily verified that $R_i = \alpha_i \bar{I} - H \oplus \Lambda$.

Then, by direct calculations with account of the imposed above assumption that $\lim\limits_{i\to\infty} \alpha_i = \infty$, it can be verified that

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

Lemma 3.2 is proven. □

4. Ergodicity condition

Theorem 4.1. (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 inequality* (1) *has an opposite sign.*

Proof. (i) As follows from [12], the sufficient condition for ergodicity of the AQTMC ξ_n, $n \geqslant 1$, is the fulfillment of the inequality

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

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

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

Calculating the vector $\mathbf{x}$ from system (3) and substituting the expression obtained into inequality (2) after some algebra we get inequality (1).

Statement (ii) of the theorem follows from (1) and the results of [12]. □

Remark 4.2. *Condition for ergodicity is easy tractable: the average arrival rate λ is less than the sum of the mean service rates in all servers of the system.*

The numerically stable algorithm for computation of vectors π_i, $i \geqslant 0$, can be found in [12].

5. Performance measures

As soon as the vectors π_i, $i \geqslant 0$, have been calculated, we are able to find various performance measures of the system.

Let us introduce the following denotations:

$\mathcal{R}_k$, $k = \overline{1,N}$, is the set of the states $\{r^{(1)},\ldots,r^{(N)}\} \in \mathcal{R}$ of the process $\{\eta_t^{(1)},\ldots,\eta_t^{(N)}\}$ such that $r^{(l)} > 0$ for $l = \overline{1,k-1}$, $r^{(k)} = 0$.

For the fixed set $(r^{(1)},\dots,r^{(N)})$, $(r^{(1)},\dots,r^{(N)})\in\mathcal{R}$, the value $l(r^{(1)},\dots,r^{(N)})$ defines the number of states having nonzero value of the components $r^{(n)}$, $n=\overline{1,N}$:

$$l(r^{(1)},\dots,r^{(N)})=\sum_{n=1}^{N}(1-\delta_{r^{(n)},0}).$$

The average number L_{orbit} of customers in the orbit is computed by

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

The probability $P_{\text{empty-orbit}}$ that the orbit is empty at an arbitrary moment is computed by

$$P_{\text{empty-orbit}}=\boldsymbol{\pi}_0\boldsymbol{e}.$$

The average number N_{busy} of busy servers at an arbitrary moment is computed by

$$N_{\text{busy}}=\sum_{i=0}^{\infty}\sum_{(r^{(1)},\dots,r^{(N)})\in\mathcal{R}}l(r^{(1)},\dots,r^{(N)})\boldsymbol{\pi}(i,r^{(1)},\dots,r^{(N)})\boldsymbol{e}.$$

The probability $P_0^{(n)}$ that the nth server, $n=\overline{1,N}$, is idle at an arbitrary moment is computed by

$$P_0^{(n)}=\sum_{i=0}^{\infty}\boldsymbol{\pi}_i(\overset{n-1}{\underset{r=1}{\otimes}}\boldsymbol{e}_{M^{(r)}+1}\otimes\mathbf{f}^{(n)}\otimes\overset{N}{\underset{r=n+1}{\otimes}}\boldsymbol{e}_{M^{(r)}+1}\otimes\boldsymbol{e}_{\bar{W}}).$$

The row vector defining the stationary probability distribution $\mathbf{p}_n$ of the status of the nth server at an arbitrary moment is given by formula

$$\mathbf{p}_n=\sum_{i=0}^{\infty}\boldsymbol{\pi}_i((\overset{n-1}{\underset{r=1}{\otimes}}\boldsymbol{e}_{M^{(r)}+1}\otimes I_{M^{(n)}+1}\otimes\overset{N}{\underset{r=n+1}{\otimes}}\boldsymbol{e}_{M^{(r)}+1})\otimes\boldsymbol{e}_{\bar{W}}),\ n=\overline{1,N}.$$

The output rate φ_n from the nth server is defined by

$$\varphi_n=\mathbf{p}_n\begin{pmatrix}0\\S_0^{(n)}\end{pmatrix},\ n=\overline{1,N}.$$

Relation $\lambda=\sum_{n=1}^{N}\varphi_n$ can be used for control of accuracy of computation of the stationary distribution of the system states.

The probability $P_0^{(\text{serv})}$ that all servers are idle at an arbitrary moment is computed by

$$P_0^{(\text{serv})}=\sum_{i=0}^{\infty}\boldsymbol{\pi}_i(\overset{N}{\underset{n=1}{\otimes}}\mathbf{f}^{(n)}\otimes\boldsymbol{e}_{\bar{W}}).$$

The probability P_{imm} that an arbitrary customer will succeed to enter the service immediately upon arrival is computed by

$$P_{\text{imm}}=\frac{1}{\lambda}\sum_{i=0}^{\infty}\boldsymbol{\pi}_i\left(\tilde{I}_\beta\otimes D_1\right)\boldsymbol{e}.$$

The share of customers, which start service immediately upon arrival by the kth server, is computed by

$$Z_k=\frac{1}{\lambda}\sum_{i=0}^{\infty}\sum_{(r^{(1)},\dots,r^{(N)})\in\mathcal{R}_k}(\boldsymbol{\pi}(i,r^{(1)},\dots,r^{(N)})\otimes D_1)\boldsymbol{e},\ k=\overline{1,N}.$$

6. Numerical results

To illustrate the feasibility and outcome of the presented algorithms as well to show the effect of correlation in arrival process, we briefly consider the following example.

Let initially the *MAP*-input be characterized by the matrices

$$D_0=\begin{pmatrix}-1.35164 & 0\\ 0 & -0.04387\end{pmatrix},$$

$$D_1 = \begin{pmatrix} 1.34265 & 0.00899 \\ 0.02443 & 0.01944 \end{pmatrix}.$$

This arrival process has the coefficient of correlation of two successive intervals between arrivals $c_{\text{cor}} = 0.2$, and the squared coefficient of variation of the intervals between customer arrivals $c_{\text{var}} = 13.4$. In the presented experiment, we will vary the average rate of the *MAP* λ what is done by multiplying the matrices D_0 and D_1 by the same scalar.

In parallel, we present the results of computation for the model where the arrival flow is defined as the stationary Poisson process with the same intensity. Let us assume that the total number N of servers be equal to 3 and $M^{(1)} = 2, M^{(2)} = 2, M^{(3)} = 3$. The retrial rates are defined by $\alpha_0 = 0,\ \alpha_i = i\alpha,\ \alpha = 1,\ i > 0$.

We will denote the *PH*-distributions of service times on three devices as $PH_1^{(\text{serv})}$, $PH_2^{(\text{serv})}$, $PH_2^{(\text{serv})}$.

$PH_1^{(\text{serv})}$ – the 2nd order hyperexponential distribution with $c_{\text{var}}^2 = 4.55$ – is characterized by the following vector and matrix:

$$\boldsymbol{\beta}^{(1)} = (0.1, 0.9), S^{(1)} = \begin{pmatrix} -2 & 0 \\ 0 & -18 \end{pmatrix}.$$

$PH_2^{(\text{serv})}$ – the 2nd order hyperexponential distribution with $c_{\text{var}}^2 = 4.54$ – is characterized by the following vector and matrix:

$$\boldsymbol{\beta}^{(2)} = (0.1, 0.9), S^{(2)} = \begin{pmatrix} -1.5 & 0 \\ 0 & -13.5 \end{pmatrix}.$$

$PH_3^{(\text{serv})}$ – the 3nd order hyperexponential distribution with $c_{\text{var}}^2 = 4.28$ – is characterized by the following vector and matrix:

$$\boldsymbol{\beta}^{(3)} = (\frac{1}{15}, \frac{2}{15}, \frac{12}{15}), S^{(3)} = \begin{pmatrix} -0.2 & 0 & 0 \\ 0 & -0.4 & \\ 0 & 0 & -2.4 \end{pmatrix}.$$

We can calculate the service rates at the corresponding servers as follows:

$$\mu_1 = 10, \mu_2 = 7.5, \mu_3 = 1$$

Let us assume that the total number N of servers be equal to 3. When we consider the service time to be exponential, assuming service rates at the corresponding servers be $\mu_1 = 10, \mu_2 = 7.5$ and $\mu_3 = 1$, correspondingly. Fig. 1 shows the behavior of the value L_{orbit} depending on the input rate λ. Fig. 2 shows the behavior of the value N_{busy} depending on the input rate λ. Fig. 3 shows the behavior of the value P_{imm} depending on the input rate λ. Fig. 4 shows the behavior of the value $P_0^{(n)}$ depending on the input rate λ under different numbers of servers. Fig. 5–7 show the behavior of the value $P_0^{(n)}$ depending on the input rate λ when $n = 1, 2, 3$, respectively, with different models.

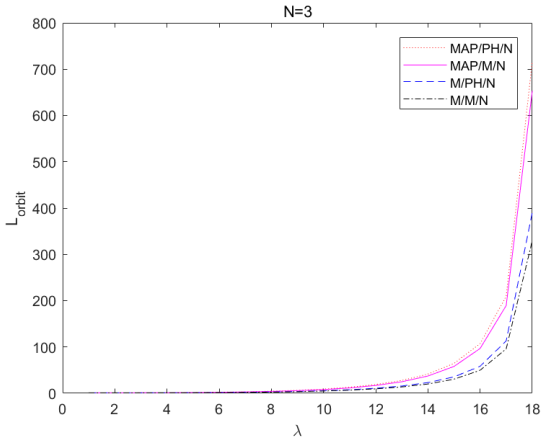

Fig. 1. Dependence of the number L_{orbit} on the input rate λ when $N = 3$ with different models

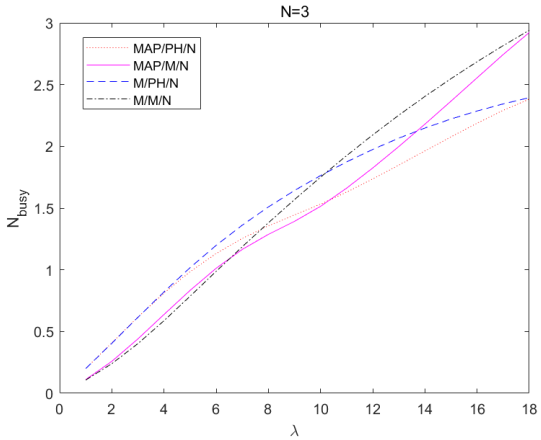

Fig. 2. Dependence of the number N_{busy} on the input rate λ when $N = 3$ with different models

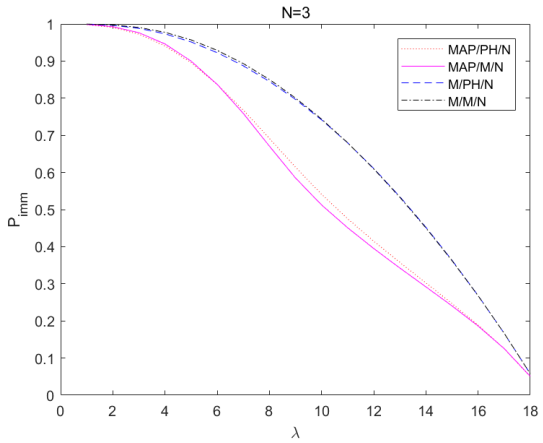

Fig. 3. Dependence of the number P_{imm} on the input rate λ when $N = 3$ with different models

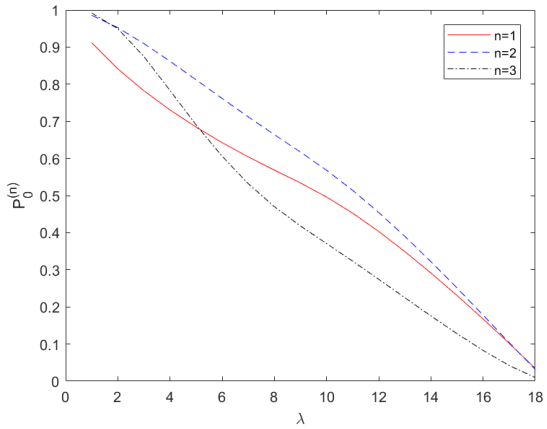

Fig. 4. Dependence of the number $P_0^{(n)}$ on the input rate λ when $N = 3$

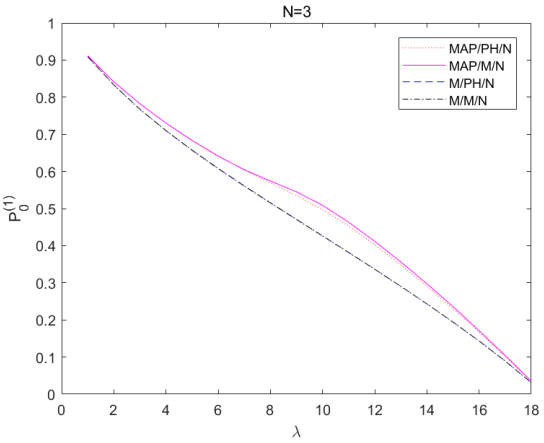

Fig. 5. Dependence of the number $P_0^{(1)}$ on the input rate λ when $N = 3$ with different models

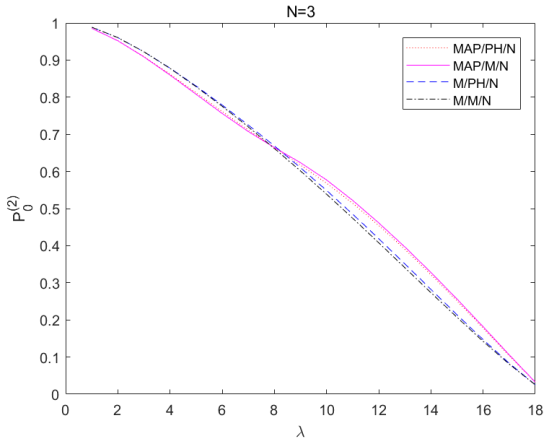

Fig. 6. Dependence of the number $P_0^{(2)}$ on the input rate λ when $N = 3$ with different models

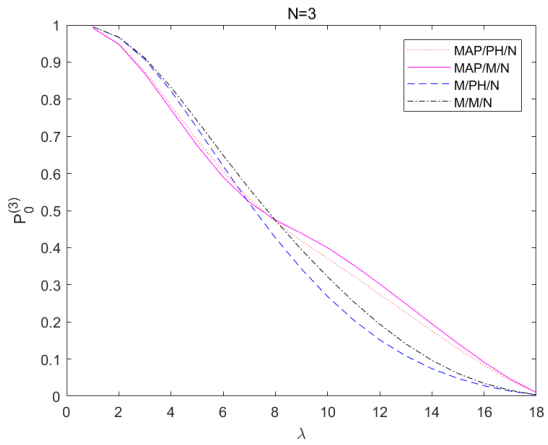

Fig. 7. Dependence of the number $P_0^{(3)}$ on the input rate λ when $N = 3$ with different models

7. Conclusion

In this paper, the algorithmic analysis of the $MAP/PH/N$ retrial queue with heterogeneous servers is presented. The obtained results are numerically illustrated in brief.

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

References

1. Falin G. I., Templeton J. G. *Retrial queues*. Routledge, New York, NY, USA, 2023.

2. Lucantoni D. New results on the single server queue with a batch Markovian arrival process. *Communication in Statistics-Stochastic Models,* 1991, vol. 7, pp. 1–46.

3. Chakravarthy S. R. The batch Markovian arrival process: A review and future work. *Adv. Probab. Theory Stoch. Process,* 2001, vol. 1, pp. 21–49.

4. Chakravarthy S. R. *Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach – Basics*. ISTE Ltd, London and John Wiley and Sons, New York, 2022.

5. Dudin A. N., Klimenok V. I., Vishnevsky V. M. *The theory of queuing systems with correlated flows*. Springer Nature, 2020.

6. Neuts M. *Matrix-geometric solutions in stochastic models*. North Chelmsford, Courier Corporation, 1994.

7. He Q. M., Li H., Zhao Y. Q. Ergodicity of the $BMAP/PH/s/s+K$ retrial queue with PH-retrial times. *Queueing Systems,* 2000, vol. 35, pp. 323–347.

8. Breuer L., Dudin A., Klimenok V. A retrial $BMAP/PH/N$ system. *Queueing Systems,* 2002, vol. 40, no. 4, pp. 433–457.

9. Liu M. Analysis of a Queue System with Repeated Calls, Heterogeneous Devices, and a Markov Arrival Process. *Informatics,* 2020, vol. 17, no. 1, pp. 48–57.

10. Liu M., Dudin A. Analysis of Retrial Queue with Heterogeneous Servers and Markovian Arrival Process. *Applied Probability and Stochastic Processes,* Springer, 2020, pp. 29–49.

11. Graham A. Kronecker products and matrix calculus with applications. *Courier Dover Publications*, Mineola, 2018.

12. 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.