Retrial queuing system with Markovian arrival flow and phase-type service time distribution

Olga Dudina a, Chesoong Kim b,⁎, Sergey Dudin a

a Belarusian State University, 4, Nezavisimosti Ave., Minsk 220030, Belarus
b Sangji University, Wonju, Kangwon 220-702, Republic of Korea

Abstract

We consider a multi-server queuing system with retrial customers to model a call center. The flow of customers is described by a Markovian arrival process (MAP). The servers are identical and independent of each other. A customer’s service time has a phase-type distribution (PH). If all servers are busy during the customer arrival epoch, the customer moves to the buffer with a probability that depends on the number of customers in the system, leaves the system forever, or goes into an orbit of infinite size. A customer in the orbit tries his (her) luck in an exponentially distributed arbitrary time. During a waiting period in the buffer, customers can be impatient and may leave the system forever or go into orbit. A special method for reducing the dimension of the system state space is used. The ergodicity condition is derived in an analytically tractable form. The stationary distribution of the system states and the main performance measures are calculated. The problem of optimal design is solved numerically. The numerical results show the importance of considering the MAP arrival process and PH service process in the performance evaluation and capacity planning of call centers.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Retrial queues take into account that a customer who cannot obtain service immediately upon arrival may return to the system after an arbitrary amount of time. Systems with retrial customers have great practical importance because they are used for the modeling and performance evaluation of telecommunication and computer networks. Retrial calls are also an integral part of call centers, and ignoring the effect of repeated attempts can lead to significant errors in predictions regarding call center operations. For background information and an overview of the present state of the art in the study of call centers, the reader is referred to the survey by Aksin, Armony, and Mehrotra (2007) and the papers by Dudin and Dudina (2011), Nah and Kim (in press), and Roubos and Jouini (2013), as well as references therein.

Queueing systems with retrial customers have been studied extensively, as described in the survey by Gomez-Corral (2006); the bibliographies by Artalejo (1999) and Artalejo (2010); the book by Falin and Templeton (1997); and the paper by Wu and Lian (2013). Retrial queues as models of call centers are described by a Markovian arrival process (MAP), the servers are identical and independent of each other. Retrial queues with impatient customers is investigated. The authors provide simple fluid and diffusion approximations to estimate the system performance measures. In the paper by Aguir et al. (2004), an M/M/N queue with retrials and balking and impatient customers is analyzed. The authors investigate this system in both the stationary regime (exact analysis using a continuous-time Markov chain and fluid approximation) and non-stationary (fluid approximation) regimes. In Shin and Choo (2009), an M/M/N queue with impatient customers and retrials is considered. The main feature is consideration of the underlying Markov chain with two countable components.

In this paper, we consider a retrial queuing system with impatient customers that can be used to model a call center. The system consists of a finite number of operators (servers) and a finite waiting line (buffer). If all operators are busy during a customer arrival epoch, the customer becomes aware of the current queue length (“visible” queue) and, based on the information provided, decides

http://dx.doi.org/10.1016/j.cie.2013.06.020

0360-8352/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
whether to wait in line or balk (leave the system). In the latter case, the customer leaves the system forever or tries his (her) luck again later. Customers are assumed to be impatient. If a customer leaves the buffer due to impatience, he (she) can also leave the system forever or make a call later.

In all papers that address the retrial queueing models of call centers, the process of customer arrival is assumed to be a Poisson process (stationary Poisson in Aguir et al. (2004), Phung-Duc (2012), Phung-Duc et al. (2013), & Shin & Choo (2009); and non-stationary Poisson in Mandelbaum et al. (2002)), and the service time distribution is assumed to be exponential (time-dependent in Mandelbaum et al. (2002)). These assumptions greatly simplify the study of systems but also reduce the accuracy of the models. Arrival flows in modern telecommunication networks and call centers in particular do not possess the properties of stationarity, memory loss, or ordinariness (see, e.g., Aktekin & Soyer (2011) and Khudaykov, Feigin, & Mandelbaum (2010)). The distribution of the service time of a customer by an operator is also not exponential (see, e.g., Fig. 3 in Khudaykov et al. (2010)). Our statistical analysis of real arrival flows and service processes of the call center of one of the largest banks in Belarus has supported these findings. In our paper, we consider a Markovian arrival process (MAP), which is more general than a Poisson arrival process and allows us to take into account the bursty nature of flows in modern call centers, and a phase type (PH) service time distribution instead of an exponential distribution. Using a MAP allows us to capture the effect of variation and correlation within the flow of customers’ calls. The first retrial queues with the MAP arrival process were considered by Dudin and Klimenok (1999) and Dudin and Klimenok (2000). The service sizes have an arbitrary distribution in Dudin and Klimenok (1999), and are controlled by a semi-Markovian process in Dudin and Klimenok (2000). However, the models considered in Dudin and Klimenok (1999) and Dudin and Klimenok (2000), have a single server. It is well known that the consideration of multi-server queues with an arbitrary distribution is an extremely difficult (if even solvable) task. Because call centers are usually modeled by multi-server queues, the approximation of the general service time distribution by the PH service time distribution is good trade-off between considering both the mean service time and the higher moments of the service time distribution in the model and the necessity to reduce the analysis of a queuing model to an investigation of a Markov process. It is known that the set of phase-type distributions is dense in the set of all positive-valued distributions; that is, it can be used to approximate any positive-valued distribution. For background information and the present state of the art in the study of multi-server retrial queues with a PH service time distribution, the reader is referred to the papers by Breuer, Dudin, and Klimenok (2002) and Dudin and Klimenok (2012). To reduce the dimension of the system state space, we use the method proposed by Ramaswami and Lucantoni for reducing the dimension of system state spaces with a phase service time distribution (see Ramaswami, 1985; Ramaswami & Lucantoni, 1985). The application of this approach allows us to calculate the performance measures of a system with a large number of servers.

The main difficulties are caused by the fact that the queuing model under study is described by a multi-dimensional Markov chain with an infinite state space and inhomogeneous state behavior. Markov chains of such a type are classified in the literature as level-dependent (or nonhomogeneous) QBD (LDQBD) processes (see, e.g., Bright & Taylor, 1995; Ramaswami, 1996). However, in our paper, we use the results for asymptotically quasi-Toeplitz Markov chains (AQTMC) (see Klímenok & Dudin, 2006) instead of LDQBD. Using the results of Klímenok and Dudin (2006), allows us to derive the ergodicity condition, whereas using LDQBD makes it impossible to obtain the constructive ergodicity condition because LDQBD does not impose any assumptions about the asymptotical behavior of the Markov chain. The ergodicity condition in Klímenok and Dudin (2006), is given in terms of a vector that satisfies a finite system of linear algebraic equations. We were able to solve this system analytically and obtain a simple and tractable ergodicity condition. An effective algorithm for computing the stationary probabilities of AQTMC is proposed. Some key performance measures are derived. The numerical results illustrating the behavior of the system characteristics depending on its parameters are presented. The problem of optimal design is solved numerically.

The rest of the paper is organized as follows. In Section 2, the mathematical model is described. The ergodicity condition and the stationary distribution of the system states are analyzed in section 3. The expressions for the main system performance measures are given in Section 4. In Section 5, the numerical results are presented. Finally, Section 6 concludes the paper.

2. Mathematical model

We consider an MAP|PH|NI/R – N queuing system with retrial customers as a model of a call center. The structure of the system is presented in Fig. 1.

The system consists of N identical independent servers (call center operators) and a buffer of capacity R – N (waiting lines). Primary customers (customers who arrive to the system for the first time) arrive according to the MAP. The MAP is defined by the underlying process \( \nu, t \geq 0 \), which is an irreducible continuous-time Markov chain with the state space \( \{0, 1, \ldots, W\} \). Arrivals occur only at the epochs of jumps in the underlying process \( \nu, t \geq 0 \). The intensities of transitions of the process \( \nu, t \geq 0 \), that are accompanied (not accompanied) by the arrival of a customer are defined by the square matrix \( D_1(\nu) \) of size \( W \times W + 1 \). The matrix \( D = D_0 + D_1 \) is an infinitesimal generator of the process \( \nu, t \geq 0 \). The stationary distribution vector \( \pi \) of this process satisfies the system of equations \( \pi D = 0, \pi e = 1 \). Here and throughout this paper \( e \) is a zero row vector and \( e \) denotes a unit column vector.

The average intensity \( \lambda \) (fundamental rate) of the MAP is defined by \( \lambda = \pi D e \). The coefficient of variation \( c_{\text{var}} \) of intervals between customer arrivals is calculated as \( c_{\text{var}} = 2 \pi (\lambda D_1 e - 1) \), and the coefficient of correlation \( c_{\text{corr}} \) of intervals between successive arrivals is given as \( c_{\text{corr}} = (\lambda D_1 e - 1)^2 / (c_{\text{var}}^2) \). Different methods for the estimation of MAP parameters using a finite set of observed data such as a set of customer arrival times recorded at a real call center, are presented, in Buchholz and Kriegel (2009), Casale, Zhang, and Smirni (2010), Horváth, Telek, and Buchholz (2005) and Ryden (2000), among other works. More information about the MAP and related research can also be found in Kim, Dudin, Dudin, and Klimenok (2012) and Lucantoni (1991).

The service time of a customer for each server has a PH distribution with an irreducible representation (P5). This can be interpreted as the time until the underlying Markov process \( \eta_t \),
\[ \tau \geq 0, \text{ with the finite state space } \{1, \ldots, M, M + 1\} \text{ reaches the single absorbing state } M + 1 \text{ conditioned on the fact that the initial state of this process is selected among the states } \{1, \ldots, M\} \text{ according to the probabilistic row vector } \mathbf{p}. \text{ The transition rates of the process } \eta_t \text{ within the set } \{1, \ldots, M\} \text{ are defined by the sub-generator } \mathcal{S}, \text{ and the transition rates into the absorbing state (which lead to service completion) are given by the entries of the column vector } \mathbf{S}_0 = -5\mathbf{e}. \text{ The mean service time is calculated by } b_1 = \mathcal{P}(-\mathbf{S})^{-1}\mathbf{e}. \text{ The squared coefficient of variation is given by } \text{cvar} = (2\mathcal{P}(-\mathbf{S})^{-1}\mathbf{e})'(b_1)^2 - 1. \text{ For more information about PH distributions, see Neuts (1981). Methods of modeling the PH service process using a set of service times obtained from real systems can be found in Asmussen, Newman, and Olsson (1996), Panchenko and Thümmler (2007) and Thümmler, Buchholz, and Telek (2006), among other works.}

If a primary customer encounters a free server upon arrival, this customer receives service immediately. If all servers are busy during an arbitrary customer arrival epoch but the buffer is not full, the customer can leave the system forever without service, enter the buffer and wait until a server becomes available, or enter a so-called orbit and try again later after an arbitrary amount of time. A customer who arrives to the system from orbit is called a retrial customer.

If all servers are busy and the buffer is full during an arbitrary (primary or retrial) customer arrival epoch, then this customer either leaves the system forever with probability \( q_1 \) or goes into orbit with the complementary probability. If all servers are busy and there are \( l, l' \in \{0, \ldots, R - N - 1\} \), customers in the buffer during an arbitrary customer arrival epoch, the arriving customer chooses not to wait in the buffer with probability \( p_1 \) or moves to the buffer with the complementary probability. In the former case, he (she) leaves the system forever with probability \( q_1 \) or goes into orbit with the complementary probability. The dependence of the probability of joining a buffer on the current number of customers in the system realizes the conception of the visible queue.

Customers are impatient, i.e., a customer leaves the buffer after an exponentially distributed time described by the parameter \( \theta_t \), \( 0 < \theta_t < \infty \), due to a lack of service. In the case of leaving the buffer due to impatience, the customer leaves the system forever with probability \( q_1 \) or goes into orbit with the complementary probability. We assume that the total flow of retrials is such that the probability of generating a retrial attempt in an interval of infinitesimal length \( \{t, t + \Delta t\} \) is equal to \( \alpha \Delta t + o(\Delta t) \) when the number of customers in the orbit is equal to \( i \), \( i > 0 \), \( q_0 = 0 \). The orbit capacity is supposed to be unlimited. We do not fix the explicit dependence of the intensity \( \alpha_t \) on \( i \) assuming only that \( \lim_{i \to \infty} \alpha_t = \infty \). Note that such dependence describes the classic retrial strategy (\( \alpha = \ell x \)) and the linear strategy (\( \alpha_t = \ell x + \gamma, x > 0 \)) as special cases.

The goal of this paper is to obtain the ergodicity condition, the stationary distribution of system states, and some key performance measures.

### 3. Process of system states

The behavior of the system under consideration can be described in terms of the regular irreducible continuous-time Markov chain \( \xi_t = \{\xi_t, \eta_t, \nu_t, \eta_t^{(m)}\}, t \geq 0, \) where \( \xi_t \) is the number of customers in the orbit, \( \eta_t = 0 \); \( \nu_t \) is the number of customers in the system (in the buffer and in service), \( \eta_t \in \{0, \ldots, R\} \); \( \nu_t \) is the state of the underlying process of the MAP, \( \nu_t \in \{0, \ldots, W\} \); and \( \eta_t^{(m)} \) is the number of servers at phase \( m \) of service, with \( \eta_t^{(m)} \in \{0, \ldots, \min(n_t, \nu_t)\} \); \( m \in \{1, \ldots, M\} \); \( \sum_{m=1}^{M} \eta_t^{(m)} = \min(n_t, \nu_t) \), during the epoch \( t, t \geq 0 \). Note that the definition of the components \( \eta_t^{(m)} \) is chosen according to the approach by Ramaswami and Lucantoni (see Ramaswami, 1985; Ramaswami & Lucantoni, 1985), which allows us to significantly reduce the dimension of the state space of the Markov chain \( \xi_t \).

**Lemma 1.** The infinitesimal generator \( Q \) of the Markov chain \( \xi_t, t \geq 0, \) has a block-tridiagonal structure:

\[
Q = \begin{pmatrix}
0 & 0 & 0 & \cdots & 0 \\
0 & Q_{1,0} & Q_{1,1} & \cdots & 0 \\
0 & Q_{2,1} & Q_{2,2} & \cdots & 0 \\
& \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & Q_{N-1,N-1}
\end{pmatrix},
\]

The non-zero blocks \( Q_{ij}, i, j \geq 0, \) have the following form:

\[
Q_{ij} = \begin{pmatrix}
C_i^{(0)} & C_i^{(1)} & \cdots & 0 \\
0 & C_i^{(2)} & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0
\end{pmatrix}, i \geq 0,
\]

\[
Q_{i-1,j} = \begin{pmatrix}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0
\end{pmatrix}, i > 1,
\]

\[
Q_{i,j} = Q^* = \begin{pmatrix}
E^{(0)} & E^{(1)} & \cdots & 0 \\
0 & E^{(2)} & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0
\end{pmatrix}, i \geq 0,
\]

where

\[
\begin{align*}
D_{i,0} & \equiv \lambda_i (N, S) + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in} + (1 - (1 - \omega_0) \mu_i) \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in}, \\
D_{i,1} & \equiv \lambda_i (N, S) + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in}, \\
D_{i,n} & \equiv \lambda_i (N, S) + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in}, \\
D_{i,R} & \equiv \lambda_i (N, S) + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in}, \quad n \in \{N + 1, \ldots, R\},
\end{align*}
\]

\[
C_i = \begin{pmatrix}
(1 - \omega_0) \lambda_i (N, S) + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in} & \lambda_i (N, S) + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in} & \cdots & \lambda_i (N, S) + \lambda_i \mathcal{M}_{in} + \lambda_i \mathcal{M}_{in}
\end{pmatrix}, \quad n \in \{1, \ldots, N\},
\]

\[
E^{(0)} = \begin{pmatrix}
\mathcal{D}_{1,0} & \mathcal{D}_{1,1} & \cdots & \mathcal{D}_{1,n}
\end{pmatrix}, \quad n \in \{0, \ldots, N - 1\},
\]

\[
E^{(1)} = \begin{pmatrix}
\mathcal{D}_{2,0} & \mathcal{D}_{2,1} & \cdots & \mathcal{D}_{2,n}
\end{pmatrix}, \quad n \in \{0, \ldots, N - 1\},
\]

\[
E^{(2)} = \begin{pmatrix}
\mathcal{D}_{3,0} & \mathcal{D}_{3,1} & \cdots & \mathcal{D}_{3,n}
\end{pmatrix}, \quad n \in \{0, \ldots, N - 1\},
\]

\[
E^{(3)} = \begin{pmatrix}
\mathcal{D}_{4,0} & \mathcal{D}_{4,1} & \cdots & \mathcal{D}_{4,n}
\end{pmatrix}, \quad n \in \{0, \ldots, N - 1\},
\]

\[
E^{(4)} = \begin{pmatrix}
\mathcal{D}_{5,0} & \mathcal{D}_{5,1} & \cdots & \mathcal{D}_{5,n}
\end{pmatrix}, \quad n = R.
\]
Here

- \( I \) is the identity matrix, and \( O \) is a zero matrix of appropriate dimension;
- \( \otimes \) and \( \circ \) indicate the Kronecker sum and product, respectively (see Graham, 1981);

\[
\tilde{S} = \begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix};
\]

\[
K_n = \begin{pmatrix} n + M - 1 \\ M - 1 \end{pmatrix}, \quad n \in \{0, \ldots, N\};
\]

\[
A^{(0)} = - \text{diag}\{A_i(N, S)\} + L_{n, i}(\tilde{N}, \tilde{S}) e, \quad i \in \{1, \ldots, N\};
\]

\[A^{(0)} = O_{1 \times 1}.
\]

A detailed description of the matrices \( P_i(\beta), i \in \{0, \ldots, N - 1\}, A_i(N, S), L_{n, i}(N, \tilde{S}), i \in \{0, \ldots, N\} \), and the algorithms for their calculation are presented in Kim, Dudin, Taramin, and Baek (2013), Ramaswami (1985) and Ramaswami and Lucantoni (1985).

**Proof.** The proof of Lemma 1 is implemented by the analysis of all transitions of the Markov chain \( \zeta_t, t \geq 0 \), during the interval of infinitesimal length and rewriting the intensities of these transitions in block matrix form.

Because no more than one arrival can occur and no more than one service can be completed in the system during an interval of an infinitesimal length, the blocks \( Q_{ij} = 0 \) if \( |i - j| > 1 \); thus, the generator \( Q \) has a block-tridiagonal structure.

The entries of the blocks \( Q_{ij}, i, j \geq 0 \), define the intensities of the transitions of the chain \( \zeta_t, t \geq 0 \), that increase the number of customers in the orbit by one. The subdiagonal blocks \( E^{(0)} \), \( n \in \{0, \ldots, R\} \), of the matrix \( Q_{ij(1)} \) define the intensities of transitions that decrease the number of customers in the system by one. This loss can occur if a customer leaves the buffer due to impatience and goes into orbit. The intensities of this event are given by entries of the matrix \( (1 - q_1) (n - N) \Gamma_{K_n}, n \in \{N + 1, \ldots, R\} \). The diagonal blocks \( E^{(1)} \), \( n \in \{0, \ldots, R\} \), define the intensities of transitions that do not change the number of customers in the system, which can only happen if a primary customer enters orbit after refusing to join the buffer or because of a buffer overflow. The intensities of the former event are given by the entries of the matrix \( (1 - q_2) p_n D_{n} \otimes I_{K_n}, n \in \{N, \ldots, R\} \), while the matrix \( (1 - q_1) D_{1} \otimes I_{K_n} \) defines the intensities of the latter event.

The entries of the blocks \( Q_{ij}, i, j \geq 1 \), define the intensities of the transitions of the chain \( \zeta_t, t \geq 0 \), that decrease the number of customers in the orbit by one. The updiagonal blocks \( \tilde{E}^{(0)} \), \( n \in \{0, \ldots, R - 1\} \), of the matrix \( Q_{ij(0)} \) define the intensities of transitions that increase the number of customers in the system by one. This can only occur if a retrial customer makes a successful attempt and is accepted into the system. The intensities of this event are given by the entries of the matrix \( \tilde{A}_i(N, \tilde{S}) \Gamma_{K_n} \), \( n \in \{0, \ldots, N - 1\} \), if there is a free server or by the entries of the matrix \( (1 - p_{n, N}) \tilde{A}_i(N, \tilde{S}) \Gamma_{K_n}, n \in \{N, \ldots, R - 1\} \), if all servers are busy but the buffer is not full. The diagonal blocks \( \tilde{E}^{(1)} \), \( n \in \{0, \ldots, R\} \), define the intensities of transitions that do not change the number of customers in the system. This can only occur if the attempt of a retrial customer to enter the system is unsuccessful and this customer leaves the orbit and the system forever. The intensities of this event are given by entries of the matrices \( q_1 p_{n, N} \tilde{A}_i(N, \tilde{S}) \), \( n \in \{N, \ldots, R - 1\} \), if the buffer is not full, and those of the matrix \( q_1 \tilde{A}_i(N, \tilde{S}) \Gamma_{K_n} \) otherwise.

The non-diagonal entries of the blocks \( Q_{ij, i \geq 0} \), define the intensities of the transitions of the chain \( \zeta_t, t \geq 0 \), that do not change the number of customers in orbit. The updiagonal blocks \( \tilde{E}^{(0)} \), \( n \in \{0, \ldots, R - 1\} \), define the intensities of transitions that increase the number of customers in the system by one. This can occur if a primary customer is accepted into the system. The intensities of this event are given by the entries of the matrices \( D_{i} \otimes P_i(\beta), n \in \{0, \ldots, N - 1\} \), if there is a free server and by those of the matrices \( (1 - p_{n, N}) \tilde{A}_i(N, \tilde{S}) \Gamma_{K_n}, n \in \{N, \ldots, R - 1\} \), if all servers are busy but the buffer is not full. The subdiagonal blocks \( E^{(0)} \), \( n \in \{1, \ldots, R\} \), define the intensities of transitions that decrease the number of customers in the system by one. This can only occur if service is completed or a customer leaves the buffer due to impatience and leaves the system forever. The intensities of the former event are given by the entries of the matrices \( I_{K_n} \otimes L_{n, i}(N, \tilde{S}), n \in \{1, \ldots, N\} \), if there is no customer in the buffer and by those of the matrices \( I_{K_n} \otimes (L_{n, i}(N, \tilde{S}) P_{n, i(1)}(\beta)), n \in \{N + 1, \ldots, R\} \), otherwise. The matrices \( q_1 (n - N) \Gamma_{K_n}, n \in \{N + 1, \ldots, R\} \), define the intensities of the latter event. The non-diagonal entries of the matrices \( \tilde{E}^{(0)} \), \( n \in \{0, \ldots, R\} \), define the intensities of transitions that do not change the number of customers in orbit of the system. The diagonal entries of the matrix \( Q_j \) are negative, and the modulus of each entry defines the total intensity of leaving the corresponding state of the chain \( \zeta_t, t \geq 0 \). \( \square \)

**Corollary 1.** The Markov chain \( \zeta_t, t \geq 0 \), belongs to the class of continuous-time asymptotically quasi-Toeplitz Markov chains (AQTMCh) (see Klimenko & Dudin 2006).

As follows from Klimenko and Dudin (2006), the sufficient condition for the ergodicity of the AQTMCh, \( \zeta_t, t \geq 0 \), is expressed in terms of the matrices

\[
Y_0 = \lim_{t \to \infty} R^{-1} Q_{i, 1-1}, \quad Y_1 = \lim_{t \to \infty} R^{-1} Q_{i, 1}, \quad Y_2 = \lim_{t \to \infty} R^{-1} Q_{i, 1, 1-1},
\]

where \( R \) is a diagonal matrix with diagonal entries defined as the moduli of the corresponding diagonal entries of the matrix \( Q_{ij}, i \geq 0 \).

In our case, it can be easily verified that \( R \) is the block diagonal matrix with the diagonal blocks \( T^{(0)}_i, n \in \{0, \ldots, R\} \), \( i \geq 0 \), defined as follows:

\[
T^{(0)}_i = \begin{pmatrix}
\begin{array}{ccc}
A & Z_{n, N} & (n - 1)! \\
\Sigma & \tilde{A}_i(N, \tilde{S}) \Gamma_{K_n} & (n - N)! \\
A & Z_{n, N} & (n - 1)! \\
\Sigma & \tilde{A}_i(N, \tilde{S}) \Gamma_{K_n} & (n - N)! \\
A & Z_{n, N} & (n - 1)! \\
\Sigma & \tilde{A}_i(N, \tilde{S}) \Gamma_{K_n} & (n - N)! \\
A & Z_{n, N} & (n - 1)! \\
\Sigma & \tilde{A}_i(N, \tilde{S}) \Gamma_{K_n} & (n - N)! \\
\end{array}
\end{pmatrix},
\]

where \( \Sigma, A \) and \( Z_n \) are diagonal matrices with diagonal entries defined by the diagonal entries of the matrices \( D_n, -D_0, \) and \( A_{ij}(N, S), n \in \{0, \ldots, N\} \), respectively. The explicit form of the square matrices \( Y_{K_n}, k = 0, 1, 2, \) of size \( \overline{W} \sum_{n=0}^{N} K_n + (R - N)K_n \) is as follows:

- If \( q_1 \neq 0 \), then

\[
Y_0 = \begin{pmatrix}
B^{(0)} & B^{(0)} & 0 & \cdots & 0 \\
0 & B^{(0)} & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & B^{(R-1)} & B^{(R-1)} \\
0 & 0 & \cdots & 0 & B^{(0)}
\end{pmatrix},
\]

\[
Y_1 = O, \quad Y_2 = O,
\]

where
\[ y^{(0)}(\mathbf{T} + \mathbf{F} + \mathbf{P} + \mathbf{e}) = y^{(0)}. \]

or, after some algebra,

\[ y^{(0)}(D \otimes I_{K_0} + I_{\mathbf{T}} \otimes (A_0(N,S) + L_0(N,S)P_{N-1}(\beta) + A^{(N)})) = 0. \]

Hence, the vector \( y^{(0)} \) has the following form:

\[ y^{(0)} = c(\chi \otimes \delta), \]

where \( \chi \) is the stationary distribution vector of the process \( v \), \( \delta \) is the stationary distribution vector of the process \( \eta = \{ \eta_1^{(1)}, \ldots, \eta_1^{(N)} \} \), which is the unique solution to the following system

\[ \delta(A_0(N,S) + L_0(N,S)P_{N-1}(\beta) + A^{(N)}) = 0. \delta \mathbf{e} = 1, \]

and \( c \) is a normalizing multiplier.

By the direct substitution of (6) into (3), it can be verified that the vector \( y^{(R-1)} \) has the following form:

\[ y^{(R-1)} = c(\chi \otimes \delta)(I_{\mathbf{T}} \otimes (L_0(N,S)P_{N-1}(\beta)) + (R - N)(\mathbf{r}_{K_0}^{\otimes}))^{(R-1)} \] 

Substituting the vectors \( y^{(0)} \) and \( y^{(R-1)} \) from (6) and (7) into inequality (5) and performing some algebra, we obtain the inequality

\[ \left( \frac{1 - p_{K_0}}{1 - (1 - q) p_{K_0}} \right)^{N - 1} \delta(A_0(N,S)e + (R - N)\mathbf{e}) > (1 - q)e + (R - N)\mathbf{e}. \]

Taking into account that \( \delta(A_0(N,S)e) = N\mu \), where \( \mu = 1/b_1 \) is the mean service rate, after some algebra, inequality (8) can be rewritten as

\[ N\mu + q(1 - (R - N)\theta) + \frac{q_{N-1} + q_{N-1}}{1 - p_{K_0}} > (1 - q) \theta. \]

Inequality (9) has an intuitively evident meaning: the Markov chain \( \xi_c \) is ergodic if the average intensity of customers' leaving the system forever (the left side of inequality (9)) exceeds the average arrival rate of primary customers (the right side of inequality (9)) under the condition that the system is overloaded. A customer leaves the system forever in three cases:

- His (her) service is completed. The value \( N\mu \) is the mean service rate of customers given that \( N \) servers are busy.
- He (she) refuses to wait in the buffer due to impatience and leaves the system forever. The value \( q_1(R - N)\theta \) defines the intensity of customer loss due to impatience.
- As a retrial customer, he (she) refuses to join the buffer. The value \( q_2(R - N)\theta \) is the intensity of buffer space release, while \( \frac{q_{N-1} + q_{N-1}}{1 - p_{K_0}} \) is the mean number of customers that are lost from orbit due to an unwillingness to join the buffer at a buffer release moment. Note that the number of lost customers has a geometric distribution.

Thus, we have derived the ergodicity condition for the AQTM C under consideration. This condition is given by the following theorem.

**Theorem 2.** If probability \( q_1 \) is not equal to 0, then the Markov chain \( \xi_c \), \( t \geq 0 \), is ergodic for any choice of system parameters.

In case \( q_1 = 0 \), the Markov chain \( \xi_c \), \( t \geq 0 \), is ergodic if inequality (9) holds true.

In what follows, we suppose that inequality (9) is fulfilled if \( q_1 = 0 \).

Denote the stationary probabilities of system states as follows:

\[ \pi(i, n, v, \eta^{(1)}, \ldots, \eta^{(M)}) = \lim_{t \to \infty} P\{i = i, n = n, v = v, \eta^{(1)} = \eta^{(1)}, \ldots, \eta^{(M)} = \eta^{(M)}\}. \]
\[ i \geq 0, \ n \in \{0, \ldots, R\}, \ v \in \{0, \ldots, W\}, \ \eta^{(m)} \in \{0, \ldots, \min\{n_i, N\}\}, \ m \in \{1, \ldots, M\}. \]

Let us form the row vectors \( \pi(i, n, v) \) of the probabilities \( \pi(i, n, v, \eta^{(1)}), \ldots, \pi(i, n, v, \eta^{(M)}) \), \( i \geq 0, \ n \in \{0, \ldots, R\}, \ v \in \{0, \ldots, W\} \), enumerated in the reverse lexicographic order of the components \( \eta^{(1)}, \ldots, \eta^{(M)} \) and the row vectors \( \pi_i \) as follows:
\[ \pi(i, n) = (\pi(i, n, 0), \pi(i, n, 1), \ldots, \pi(i, n, W)), \ n \in \{0, \ldots, R\}, \pi_i = (\pi(i, 0), \pi(i, 1), \ldots, \pi(i, R)), \ i \geq 0. \]

It is well known that the probability vectors \( \pi_i, i \geq 0 \), satisfy the following system of linear algebraic equations:
\[ (\pi_0, \pi_1, \ldots)Q = 0. \quad (\pi_0, \pi_1, \ldots)e = 1 \tag{10} \]
where \( Q \) is the infinitesimal generator of the Markov chain \( \zeta_t, t \geq 0 \).

Taking into account that the matrix \( Q \) has a block tridiagonal structure, to solve system (10), we can propose the following numerically stable algorithm.

**Theorem 3.** The vectors \( \pi_i, i \geq 0 \), are calculated as \( \pi_i = \pi_0 F_i, i \geq 1 \),
where the matrices \( F_i \) are calculated using the recurrent formulas
\[ F_0 = I, \ F_i = -F_{i-1}Q^{-1}(Q_{i-1} + Q^-G_i)^{-1}, i \geq 1, \]
the matrices \( G_i \) are computed using the backward recursion
\[ G_i = -(Q_{i-1, i} + Q^- G_{i-1})^{-1} Q_{i-1, i}, i \geq 0. \]
and the vector \( \pi_0 \) is the unique solution to the system
\[ \pi_0(Q_{0,0} + Q^- G_0) = 0, \ \pi_0 \sum_{i=1}^{\infty} F_i e = 1. \]

Proof follows from Klimenok and Dudin (2006), taking into account the structure of the generator \( Q \).

Note that recursion (12) is backward and, to compute the matrix \( G_0 \), we must know all matrices \( G_i, i \geq 1 \).

As \( i \) approaches infinity, the sequence of the matrices \( G_i \) tends to the matrix \( G \) that is the minimal non-negative solution to the matrix equation (see Klimenok & Dudin (2006))
\[ G = Y_0 + Y_1 G + Y_2 G^2. \tag{13} \]
This implies that, for any predefined small number \( \varepsilon_c > 0 \), there exists such a value \( \varepsilon_c \) that the norm of the matrix \( G - G \) is less than \( \varepsilon_c \) for all \( i, i \geq i_0 \). If \( \varepsilon_1 = 0 \), the matrix \( G \) is equal to the matrix \( Y_0 \).
Otherwise, matrix Eq. (13) can be solved, e.g., using the following iteration method:
\[ G^{(0)} = I, \ G^{(m+1)} = (I - Y_1)^{-1}(Y_0 + Y_2 G^{(m)})^2, m > 0. \]
We compute the matrix \( G \) until the norm of the matrix \( G^{(m+1)} - G^{(m)} \) becomes less than a small pre-assigned value. When the dimension of the matrix \( G \) is large, this iterative procedure can be time-consuming, and the logarithmic reduction method of Latouche and Ramaswami (1999) can be used to more quickly compute the matrix \( G \). All the matrices \( F_i, i \geq 0 \), have non-negative entries and, because we assume that the chain under consideration is ergodic, the norm of the matrix \( F_i \) tends to zero as \( i \) approaches infinity. Therefore, the recursive calculation of the matrices \( F_i \) from (11) can be halted when the norm of the matrix \( F_i \) becomes less than some pre-assigned accuracy level \( \varepsilon_i \). Thus, the problem of computing the vectors \( \pi_i, i \geq 0 \), can be considered solved.

4. Performance measures

The average number of customers in the orbit is \( L_{\text{orb}} = \sum_{i=1}^{\infty} i \pi_i e \).

The average number of customers in the system is \( L_{\text{sys}} = \sum_{n=0}^{\infty} \sum_{i=1}^{\infty} n \pi(i, n) e \).

The average number of customers in the buffer is \( N_{\text{buf}} = \sum_{n=0}^{\infty} \sum_{i=1}^{\infty} n (n - N) \pi(i, n) e \).

The average number of busy servers is \( N_{\text{bus}} = \sum_{n=0}^{\infty} \sum_{i=1}^{\infty} \min(n, N) \pi(i, n) e \).

The intensity of flow of customers that receive service in the system is
\[ \lambda_{\text{ser}} = N_{\text{buf}}/D_1 = \sum_{i=0}^{\infty} \sum_{n=1}^{R} \pi(i, n) D_1 ÷ L_{\text{max}}(0, N, n) (N, S) e. \]

The probability that an arbitrary primary customer is rejected at the entrance to the system due to buffer overflow and goes into orbit or leaves the system forever is
\[ p_{1}^{\text{int-los}} = \lambda^{-1} \sum_{i=0}^{\infty} \pi(i, R) D_1 ÷ L_{\text{sys}} e. \]

The probability that an arbitrary retrial customer is rejected at the entrance to the system due to buffer overflow and goes into orbit or leaves the system forever is
\[ p_{1}^{\text{int-los}} = x^{-1} \sum_{i=1}^{\infty} x_1 \pi(i, R) e \]
where \( x = \sum_{i=1}^{\infty} x_{i-1} \pi(i, n) e \).

The probability that an arbitrary primary customer arrives when all servers are busy, the buffer is not full, and the customer does not join the buffer and leaves the system forever or goes into orbit is given as
\[ p_{1}^{\text{los-los}} = \lambda^{-1} \sum_{i=0}^{\infty} \sum_{n=0}^{R-1} \pi(n, i) D_1 ÷ L_{\text{sys}} e. \]

The probability that an arbitrary retrial customer arrives when all servers are busy, the buffer is not full, and the customer does not join the buffer and leaves the system forever or goes into orbit is given as
\[ p_{1}^{\text{los-los}} = x^{-1} \sum_{i=1}^{\infty} \sum_{n=0}^{R-1} \pi(n, i) e. \]

The probability that, after arrival, an arbitrary primary customer will go to the buffer and leave it due to impatience is calculated as
\[ p_{1}^{\text{imp-los}} = \lambda^{-1} \sum_{i=0}^{\infty} \sum_{n=0}^{R-1} (1 - p_{n-1}) \pi(n, i) D_1 ÷ L_{\text{sys}} (e_{n-1} - z(n - N + 1)) \]
where the column vectors \( z(n) \) consisting of the probabilities \( z(n, \eta^{(1)}), \ldots, \eta^{(M)} \) enumerated in the reverse lexicographic order of the components \( \eta^{(1)}, \ldots, \eta^{(M)} \) are given as follows:
\[ z(1) = (\delta \delta A^{-1} Le, z(1) = (n \delta A^{-1} L + (n - 1) \delta) z(n - 1), \]
\[ n \in \{2, \ldots, R - N\} \]
where \( A = A_0(N, S) + A^T, L = L_0(N, S) p_{N-1}(B) \).

Note that \( z(n, \eta^{(1)}), \ldots, \eta^{(M)} \) defines the probability, that during the wait time in the buffer, the customer does not leave the system due to impatience conditioned on the fact that during his (her) arrival epoch there are \( n - 1 \) customers in the buffer and the states of the processes \( \eta^{(1)}, \ldots, \eta^{(M)} \) are \( \eta^{(1)}, \ldots, \eta^{(M)} \), respectively.

The probability that an arbitrary retrial customer will go to the buffer and leave it due to impatience is calculated as
\[ p_{1}^{\text{imp-los}} = x^{-1} \sum_{i=1}^{\infty} \sum_{n=0}^{R-1} (1 - p_{n-1}) x_1 \pi(n, i) (e_{n-1} ÷ L_{\text{sys}}) (e_{n-1} - z(n - N + 1)). \]
The probability that an arbitrary primary customer will go into orbit is calculated as 
\[ p_{\text{in-orbit}} = (1 - q_1)q_1^{\text{pre-loss}} + (1 - q_2)q_1^{\text{pre-cas}} + (1 - q_3)q_1^{\text{imp-loss}}. \]

The probability that an arbitrary primary customer will be lost and leave the system forever is given as 
\[ p_{\text{loss}} = q_1^{\text{pre-loss}} + q_2^{\text{pre-cas}} + q_3^{\text{imp-loss}}. \]

The probability that an arbitrary primary customer will be served (without visiting the orbit) is calculated as 
\[ p_{\text{pre}} = 1 - p_{\text{in-orbit}} - p_{\text{loss}}. \]

The probability that an arbitrary retrial customer will be lost and leave the system forever is given as 
\[ p_{\text{pre}} = q_1^{\text{pre-loss}} + q_2^{\text{pre-cas}} + q_3^{\text{imp-loss}}. \]

The probability that an arbitrary retry customer will return to orbit is given as 
\[ p_{\text{return}} = (1 - q_1)p_{\text{pre-loss}} + (1 - q_2)p_{\text{pre-cas}} + (1 - q_3)p_{\text{pre-cas}}. \]

The probability that an arbitrary retry customer will be served without returning to orbit is computed by 
\[ p_{\text{return}} = 1 - p_2 - p_{\text{return}}. \]

The probability that an arbitrary primary customer accesses the server upon arrival is calculated as 
\[ p_{\text{access}} = x^{n-1} \sum_{i=0}^{n-1} \pi(i, n)(D_1 \otimes I_{K_n})e. \]

Let \( V(x) \) be the distribution function of the sojourn time of an arbitrary primary customer in the system under study and \( V(s) = \int_0^s e^{-sv} dv(x) \). \( s > 0 \), be its Laplace-Stieltjes transform (LST).

**Theorem 4.** The LST \( V(s) \) is calculated as follows:
\[ V(s) = \frac{1}{P_{X=1}} \sum_{i=0}^{\infty} \sum_{n=0}^{N-1} \pi(i, n)(D_1 \otimes I_{K_n})e \beta(s - S)^{-1}S_0 + \sum_{n=N}^{N-1} (1 - p_{n,n})\pi(i, n)(D_1e \otimes I_{K_n})V(s, n + 1) \]
where the vectors \( V(s, n), n \in \{1, \ldots, R\} \), are defined by
\[ V(s, 1) = ((s + 0)I - A)^{-1} \text{Le}(s - S)^{-1}S_0, \]
\[ V(s, n) = ((s + n0)I - A)^{-1}(L + (n - 1)0I)V(s, n - 1), \quad n \in \{2, \ldots, R\}. \]

**Proof.** Let us tag an arbitrary primary customer and keep track of his (her) service process in the system. We will derive the expression for the LST \( V(s) \) using the method of collective marks (method of additional events, method of catastrophes) (see, e.g., Kesten & Runnungenburg, 1956; van Danzig, 1955). To this end, we interpret the variable \( s \) as the intensity of some virtual stationary Poisson flow of catastrophes. Thus, \( V(s) \) has the meaning of the probability that no catastrophe arrives during the sojourn time of an arbitrary primary customer who does not visit the orbit.

If there is an idle server during the arrival epoch of the tagged customer, then this customer immediately receives service. The probability of this event is \( \frac{1}{P_{X=1}} \sum_{i=0}^{N-1} \pi(i, n)(D_1 \otimes I_{K_n})e \). In this case, the probability that no catastrophe arrives during the sojourn time is equal to the probability that no catastrophe arrives during the service time and is defined as \( \beta(s - S)^{-1}S_0 \).

If all servers are busy during the arrival epoch of the tagged customer, he (she) joins the buffer. Let \( \tau(s, n, n^{(1)}, \ldots, n^{(M)}) \) be the LST of the distribution of the tagged primary customer’s sojourn time conditioned on the fact that, at the given moment, the position of the tagged customer in the buffer is equal to \( n, \quad n \in \{1, \ldots, R\} \), and the states of the processes \( n^{(1)}, \ldots, n^{(M)} \), \( t > 0 \), are \( n^{(1)}, \ldots, n^{(M)} \), respectively. Let us form the vectors \( V(s, n) \) of these LSTs enumerated in the reverse lexicographic order of the components \( n^{(1)}, \ldots, n^{(M)} \). If we find the vectors \( V(s, n) \), then the probability that the tagged customer will be admitted to the system during the epoch when all servers are busy and no catastrophe arrives during its sojourn time can be written as
\[ \frac{1}{P_{X=2}} \sum_{i=0}^{N-1} \sum_{n=0}^{R-1} (1 - p_{n,n})\pi(i, n)(D_1e \otimes I_{K_n})V(s, n + 1). \]

Based on a probabilistic sense of the LST, we obtain the following system for calculation of the vectors \( V(s, n) \):
\[ (-s + 0)\beta(s, n) + (1 - \delta_{n,1})(L + (n - 1)0I)V(s, n - 1) + \delta_{n,1}\text{Le}(s - S)^{-1}S_0 = 0, \quad n \in \{1, 2, \ldots, R\}. \]

where \( \delta_{ij} \) indicates the Kronecker delta.

It is easy to verify that the solution of (16) can be written as (14) and (15). \( \square \)

**Theorem 5.** The LST \( w(s) \) of the distribution of the sojourn time of an arbitrary serviced retrial customer at his (her) last visit to the system is calculated as follows:
\[ w(s) = \frac{1}{P_{X=2}} \sum_{i=0}^{\infty} \sum_{n=0}^{N-1} \pi(i, n)\beta(s - S)^{-1}S_0 + \sum_{n=0}^{R-1} (1 - p_{n,n})\pi(i, n)(e_0 \otimes I_{K_n})V(s, n + 1) \]
This proof is analogous to the previous proof.

**Corollary 2.** The average sojourn time of an arbitrary serviced primary customer in the system is defined as
\[ V_1^{(1)} = \frac{1}{P_{X=2}} \sum_{i=0}^{\infty} \sum_{n=0}^{N-1} b_1 \pi(i, n)(D_1e \otimes I_{K_n})e \]
\[ - \sum_{n=0}^{R-1} (1 - p_{n,n})\pi(i, n)(e_0 \otimes I_{K_n})V(0, n + 1) \]
where
\[ V_1^{(1)}(0, 1) = -(0I - A)^{-1}[V_1^{(1)}(0, 1) + b_1\text{Le}[n]], \]
\[ V_1^{(1)}(0, n) = -(n0I - A)^{-1}[V_1^{(1)}(0, n) - (L + (n - 1)0I)V_1^{(1)}(0, n - 1)], \quad n \in \{2, \ldots, R\}. \]

**Proof.** Formula (17) is based on the definition \( V_1^{(1)} = -V(s)|_{s=0}. \) \( \square \)

**Corollary 3.** The average sojourn time of an arbitrary serviced retrial customer in the system is defined as
\[ V_2^{(1)} = -w(s)|_{s=0} = \frac{1}{P_{X=2}} \sum_{i=0}^{\infty} \sum_{n=0}^{N-1} b_1 \pi(i, n)e \]
\[ - \sum_{n=0}^{R-1} (1 - p_{n,n})\pi(i, n)(e_0 \otimes I_{K_n})V(0, n + 1) \]
5. Numerical examples

To demonstrate the feasibility of the developed algorithms, prove the importance of the obtained results, and numerically illustrate some interesting features of the system under consideration, we present the results of three numerical experiments.

In Experiment 1, we demonstrate the dependence of system performance measures on the number of servers and investigate the impact of the coefficient of correlation and variation coefficient in the arrival flow on system operation. For these purposes, let us introduce four MAPs with the same average arrival rate \( \lambda = 8 \) but different coefficients of correlation and variation.

- The first process, coded as \( M \), is defined by the matrices
  \[
  D_0 = \begin{pmatrix} -8 & 0.245195 \\ 0.867614 & -0.867614 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 3.2595330 & 1.622061 \\ 0 & 0 \end{pmatrix}.
  \]
  It has \( c_{cor} = 0 \) and \( c_{var} = 1 \). This arrival process is a stationary Poisson process.

- The second process, coded as \( IPP \) (interrupted Poisson process), is defined by the matrices
  \[
  D_0 = \begin{pmatrix} -25.462586 & 0.245195 \\ 0.867614 & -0.867614 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 23.595330 & 1.622061 \\ 0 & 0 \end{pmatrix}.
  \]
  It has \( c_{cor} = 0 \) and \( c_{var} = 12.4 \).

- The third process, \( MAP^{0.2} \), has \( c_{cor} = 0.2 \) and \( c_{var} = 12.3 \). It is defined by the matrices
  \[
  D_0 = \begin{pmatrix} -10.813124 & 0 \\ 0 & -0.350961 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 10.741232 & 0.071892 \\ 0.195478 & 0.155483 \end{pmatrix}.
  \]

- The fourth process, \( MAP^{0.4} \), has \( c_{cor} = 0.4 \) and \( c_{var} = 12.3 \). It is defined by the matrices
  \[
  D_0 = \begin{pmatrix} -27.181500 & 0 \\ 0.008118 & -0.881779 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 26.898385 & 0.283115 \\ 0.097064 & 0.776597 \end{pmatrix}.
  \]

Note that the processes \( M \) and \( IPP \) have the same coefficient of correlation but different coefficients of variation, while the processes \( MAP^{0.2} \) and \( MAP^{0.4} \) have the same coefficient of variation but different coefficients of correlation.

The PH service process was modeled by statistical analysis of the service time intervals (including talk time and after-call work time) at the call center of a real bank. It is characterized by the vector \( \beta = (0.2, 0.8) \) and the matrix
  \[
  S = \begin{pmatrix} -0.4168 & 0.40768 \\ 0.42608 & -1.71809 \end{pmatrix}.
  \]

The mean service time \( b_1 \) is equal to 1.98, and the coefficient of variation is \( c_{var} = 2.027 \). We assume that the buffer capacity is equal to 20, the probabilities are \( q_1 = 0.4, q_2 = 0.3, q_3 = 0.2, q_4 = 0.3 \), and the balking probabilities are \( p_l = 0.02(l + 1), l \in \{0, \ldots, 19\} \), and we consider the classical retrial strategy \( z = 0, \ldots, 19 \), with \( z = 0.5 \). Let us vary the number of servers \( N \) in the interval \([1, 20]\).

Fig. 2 illustrates the dependence of the average number of customers in the orbit \( L_{orb} \), the average intensity of output flow of serviced customers \( \lambda_{out} \), the loss probability of an arbitrary primary customer \( P_{1-\text{loss}} \), and an arbitrary retrial customer \( P_{2-\text{loss}} \) at the entrance to the system due to buffer overflow on the number of servers \( N \) for the four MAP arrival processes presented above.

The evident conclusion from Fig. 2 is that an increase in the number of servers improves the system performance: the average
number of customers in orbit and the loss probabilities of customers at the entrance to the system decrease, while the flow intensity of customers that receive service in the system increases. Additionally, one can conclude from Fig. 2 that the coefficient of correlation and the coefficient of variation in the arrival process have a profound impact on the system performance measures. As the coefficient of correlation and coefficient of variation increase, the flow intensity of customers who receive service decreases, while the loss probabilities and the average number of customers in orbit increase. Fig. 2 shows that for the stationary Poisson arrival process \( (M) \), 20 servers are sufficient for good system performance; if we fix \( N = 20 \), the average number of customers in orbit tends to zero and the average intensity of output flow tends to the average arrival intensity. However, this conclusion does not hold for the other arrival flows under consideration, especially for MAP\(^{0.4}\).

Let us more closely consider the dependencies of the loss probability of an arbitrary primary customer \( P_{1}^{\text{ent-loss}} \) and an arbitrary retrial customer \( P_{2}^{\text{ent-loss}} \) at the entrance to the system due to buffer overflow on the number of servers. One can see that the behaviors of the curves that correspond to \( P_{1}^{\text{ent-loss}} \) and \( P_{2}^{\text{ent-loss}} \) are significantly different. For all arrival processes, except the stationary Poisson arrival flow, the loss probability \( P_{2}^{\text{ent-loss}} \) is significantly less than the loss probability \( P_{1}^{\text{ent-loss}} \) for a small number of servers. Therefore, for small \( N \), the probability of losing a primary customer is higher than that of losing a retrial customer. This finding can be explained as follows. When the number of \( N \) is small, the system is overloaded. For the IPP and MAP arrival processes with positive coefficient of correlation, the time intervals in which customers arrive in the system often alternate with intervals in which few customers arrive in the system. Thus, when primary customers arrive intensively, the buffer quickly becomes full and many arriving customers are lost. At the same time, in intervals when few primary customers arrive, retrial customers are better able to enter the system. For this reason, when the number of servers \( N \) is small, the loss probability \( P_{2}^{\text{ent-loss}} \) is less than the loss probability \( P_{1}^{\text{ent-loss}} \), and the loss probability \( P_{2}^{\text{ent-loss}} \) is less for the interrupted Poisson arrival flow (IPP) than for the stationary Poisson flow.

With growth of \( N \), the loss probability \( P_{1}^{\text{ent-loss}} \) decreases more quickly than the loss probability \( P_{2}^{\text{ent-loss}} \). As we can see from Fig. 2, for \( N = 20 \), the probability of losing a primary customer is less than that of losing a retrial customer at the entrance to the system. For example, if \( N = 20 \), for the MAP\(^{0.4}\) arrival process, the loss probability \( P_{2}^{\text{ent-loss}} = 0.2697 \) and \( P_{1}^{\text{ent-loss}} = 0.1989 \). This finding can be explained as follows. When \( N \) is large, e.g., \( N = 20 \), the system is not overloaded, but the system load is not uniform. The intervals when the system is overloaded alternate with intervals when the system is idle. If the customer arrives at the overloaded system and goes into orbit, he (she) may attempt to enter the system again before the system load decreases and is more likely to be rejected again than a primary customer who enters the system at an arbitrary epoch.

Fig. 3 illustrates the dependence of the loss probability of an arbitrary primary customer \( P_{1}^{\text{esc-loss}} \) and an arbitrary retrial customer \( P_{2}^{\text{esc-loss}} \) due to the unwillingness of the customer to wait and the loss probability of an arbitrary primary customer \( P_{1}^{\text{imp-loss}} \) and an arbitrary retrial customer \( P_{2}^{\text{imp-loss}} \) due to impatience on the number of servers \( N \) for different MAP arrival processes.

From Fig. 3, one can also conclude that the coefficients of correlation and variation affect the presented system performance measures and that the dependencies of performance measures that relate to primary customers are different from the dependencies of performance measures that relate to retrial customers. It is
interesting that the probabilities $P_{r}^{\text{ec-loss}}$, $r = 1.2$, increase for MAP$^{0.4}$, whereas for other arrival flows, these probabilities first increase and then decrease with increasing $N$. This finding can be explained as follows. As we can see from Fig. 3, with increasing $N$, the loss probabilities $P_{r}^{\text{ec-loss}}$, $r = 1.2$, decrease. Many customers arrive to the system when the buffer is not full but the system queue is long, and the customers decide not to wait for service and leave the system. So, $P_{r}^{\text{ec-loss}}$, $r = 1.2$, first increases due to the decrease in $P_{r}^{\text{crit-loss}}$, $r = 1.2$. Furthermore, with increasing $N$, the queue length decreases and the probabilities $P_{r}^{\text{ec-loss}}$, $r = 1.2$, begin to decrease. The fact that the probabilities $P_{r}^{\text{ec-loss}}$, $r = 1.2$, increase for MAP$^{0.4}$ if $N < 20$ can be explained by the fact that, even for $N = 20$, the system queue is still long and $P_{r}^{\text{ec-loss}}$, $r = 1.2$, will decrease if we continue to increase in the number of servers.

In Fig. 4, the dependencies of the loss probabilities $P_{r}^{\text{loss}}$, $r = 1.2$, the probability that a primary customer will be served without visiting orbit $P_{r}^{\text{v-orb}}$, and the probability that a retrial customer will be served without returning to orbit $P_{r}^{\text{v-orb}}$ on the number of servers $N$ are presented for different MAP arrival processes.

Fig. 4 shows that an increase in the number of servers $N$ decreases the loss probabilities $P_{r}^{\text{loss}}$, $r = 1.2$, and increases the probabilities $P_{r}^{\text{v-orb}}$, $r = 1.2$. The probability $P_{l}^{\text{loss}}$ is less than the probability $P_{r}^{\text{loss}}$ for the reasons mentioned above, especially for large $N$. One can also see that the probability of a primary customer successfully receiving service is essentially greater than the same probability for a retrial customer.

The dependence of the probability of the immediate access of a primary customer to the server $P_{r}^{\text{imm}}$, the probability that an arbitrary primary customer will go into orbit $P_{r}^{\text{imm-orb}}$, the average sojourn time of an arbitrary serviced primary customer $V_{1}^{\text{soj}}$, and the average sojourn time of an arbitrary serviced retrial customer $V_{2}^{\text{soj}}$ on the number of servers $N$ is presented in Fig. 5.

The probability of a primary customer receiving immediate access to the server $P_{r}^{\text{imm}}$ increases with the number of servers $N$. It is surprising that for $N < 14$, the probability $P_{r}^{\text{imm}}$ is greater for MAP arrival processes, with large coefficients of correlation and variation, than for the stationary Poisson arrival process. This finding can be explained as follows. Customers from the stationary Poisson arrival process arrive in the system with the same intensity in each time interval. For this reason, the system has a more balanced load, and it is unlikely (especially for $N < 12$) that an arbitrary primary customer arrives to the system during the epoch in which there is a free server. In MAP process with large coefficients of correlation and variation, the instantaneous arrival rate of customers changes. Thus, customers who arrive during the time intervals with low arrival intensity are more likely to enter the system when there is a free server. Correspondingly, customers who arrive in the system during the time intervals with high arrival intensity are more likely to be lost or go into orbit. Thus, the probability $P_{r}^{\text{v-orb}}$ for $N > 10$ is greater for MAP arrival flows than for the stationary Poisson flow. Because customers from MAP$^{0.4}$ are more likely to receive service immediately upon arrival than customers from other flows for $N < 13$, the average sojourn time of an arbitrary serviced primary customer $V_{1}^{\text{soj}}$ is less for MAP$^{0.4}$ than for other flows. The fact that the value of $V_{1}^{\text{soj}}$ for small $N$ is less for MAP$^{0.4}$ than for other arrival flows can be explained the same way.

In Experiment 2, we show the impact of the coefficient of variation in the service process on the main system performance measures. To this end, we consider three PH service processes with the same mean service time $b_{1} = 2.027$ but different values of the coefficient of variation $c_{\text{var}}$. 

![Fig. 4. Dependence of the loss probabilities $P_{r}^{\text{loss}}$, $r = 1.2$, and probabilities $P_{r}^{\text{v-orb}}$, $r = 1.2$, on the number of servers $N$.](image-url)
Fig. 5. Dependence of the probabilities $P_{1}^{\text{in}}$ and $P_{1}^{\text{out}}$ and the average sojourn times $V_{1}^{\text{in}}$ and $V_{1}^{\text{out}}$ on the number of servers $N$.

Fig. 6. Dependence of intensity $\lambda_{\text{out}}$, the probabilities $P_{\text{return}}$ and $P_{\text{esc}}$, and the average sojourn time $V_{1}^{\text{out}}$ on the retrial rate $\alpha$ for the service process with different coefficients of variation.
The first process corresponds to the exponential distribution (M) and is defined by the vector $\beta = (1)$ and the matrix $S = (-0.49333, 0.1)$. It has $c_{\text{var}} = 1$.

The second process has $c_{\text{var}} = 0.5$ and corresponds to the Erlang distribution with the shape parameter $k = 2$. It is defined by the vector $\beta = (1, 0)$ and the matrix $S = (0.98668, -0.98668, 0.1)$. The second process has $c_{\text{var}} = 0.5$ and corresponds to a hyper-exponential distribution of order 2. It is characterized by the vector $\beta = (0.1, 0.9)$ and the matrix $S = (-0.11504, 0.00573, 0.06901, -1.25403)$.

We assume that the arrivals are defined by the third MAP process presented in the first experiment (the arrival rate $\lambda = 8$, $c_{\text{cor}} = 0.2$ and $c_{\text{var}} = 12.4$). We fix $N = 10$ and consider the classical retrial strategy $\alpha_i = i$, $i \geq 0$, where the retrial rate $\alpha$ is varied in the interval $[0.01, 10]$. The rest of the parameters are the same as those presented in the previous experiment.

Figs. 6 and 7 show the intensity of output flow $\lambda_{\text{out}}$, the probabilities $p_{\text{imm}}$, $p_{\text{cas}}$, $p_{\text{imm}}$, $p_{\text{cas}}$, $p_{\text{out}}$, and $p_{\text{out}}$, and the average sojourn time of an arbitrary serviced primary customer in the system $V_0$ as functions of the retrial rate $\alpha$ for the three service processes introduced above.

One can conclude from Figs. 6 and 7 that the coefficient of variation in the service process affects the system performance measures. As the coefficient of variation increases, the intensity of output flow $\lambda_{\text{out}}$, the probability of immediate access $p_{\text{imm}}$, the probability of successful service of a primary customer $p_{\text{cas}}$, and the probability that an arbitrary retrial customer will return to orbit also increase. At the same time, the increase in the coefficient of variation leads to a decrease in the average sojourn time of an arbitrary primary serviced customer $V_0$ and the probabilities $p_{\text{imm}}$, $p_{\text{cas}}$, $p_{\text{out}}$, and $p_{\text{out}}$. Thus, as the coefficient of variation increases, the system performance measures related to the primary customers improve, while those related to the retrial customers worsen. The same effect occurs when the intensity of retrials increases. This finding can be explained as follows. As mentioned above, if the customer arrives at an overloaded system and goes into orbit, he (she) attempts to enter the system again. If the customer quickly makes repeated attempt (with an increase in the intensity $\alpha$, the average speed of making repeated calls by a customer also increases), the system will most likely still be overloaded, and the customer will be rejected or return to orbit again. In contrast, if this customer makes another call after a long time instead of rapidly, he (she) is more likely to receive service. If the loss probability of retrial customers increases with $\alpha$, then under a high value of $\alpha$, the primary customers experience better conditions, and therefore the system performance measures that relate to the primary customers improve.

Experiment 3. The issue of optimal choice of the number of operators is very important for effective organization of operation of a call center. In this experiment, we numerically solve the problem of optimal choice of the number of servers $N$. The aim of optimization is the maximization of a cost criterion (average profit of the system per unit time):

$$J(N) = a\lambda_{\text{out}} - b(\lambda - \lambda_{\text{out}}) - c_{\text{out}} - dN.$$  

Here, $a$ is the average profit obtained by the system by servicing one customer, $\lambda_{\text{out}}$ is the intensity of the flow of customers who successfully receive service in the system, $b$ is the cost of the loss
of one customer, \(c\) is the cost of one customer staying in orbit per unit time (holding cost), and \(d\) is the cost of maintaining one server per unit time. Our goal is to find the optimal value \(N^*\) that provides the maximum value of this cost criterion. It is worth noting, that the problem of the suitable choice of cost coefficients (in our case, the coefficients \(a, b, c, d\)) in the cost criterion always plays a crucial role in the successful implementation of optimization. We assume here that in our model, the cost coefficients are obtained from experts in the call center to which the model will be applied. Therefore, we fix the following cost coefficients in the cost criterion: \(a = 12, b = 1, c = 0.5, d = 3\). The rest of the parameters are the same as those presented in the first experiment.

The dependence of the cost criterion \(J\) on the number of servers \(N\) for different MAPs is presented in Fig. 8.

The optimal values of the cost criterion \(J(N^*)\) and the number of servers \(N^*\) for different MAPs are given in Table 1. Based on Table 1, one can conclude that the optimal number of servers \(N^*\) and values of the cost criterion \(J(N^*)\) are sensitive to the coefficient of correlation in the arrival process. For example, if we assume that the arrival flow of customers to the call center is described by a stationary Poisson arrival process, when arrival flow is actually the correlated flow MAP, we would hire \(N = 17\) operators and expect a profit \(J = 38.5986\) but receive a profit of only \(4.53273\), which is more than eight times less than the expected value.

Based on these numerical experiments, the coefficients of variation and correlation in the arrival flow and variation coefficient in the service time distribution strongly affect the system performance measures. If the arrival flow in the real call center is modeled as a Poisson arrival flow and the service time distribution as exponential, the fact that the arrival flow is not Poisson and the service time is not exponential distribution can induce very large errors in the prediction of system performance measures. The results confirm the importance of taking into account the coefficients of correlation and variation in the arrival process and the variation of the service time distribution for performance evaluation and the correct prediction of the system operation.

### Table 1

<table>
<thead>
<tr>
<th>MAP</th>
<th>(J(N^*))</th>
<th>(N^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPP</td>
<td>27.9771</td>
<td>17</td>
</tr>
<tr>
<td>MAP(^{0.4})</td>
<td>28.5806</td>
<td>19</td>
</tr>
<tr>
<td>MAP(^{0.4})</td>
<td>4.69819</td>
<td>16</td>
</tr>
</tbody>
</table>

### References


