Optimization of multi-channel queuing systems with a single retail attempt: Economic approach

Oksana Pryshchepa\textsuperscript{a}, Oksana Kardasha\textsuperscript{b}, Alina Yakymchuk\textsuperscript{a}, Mykola Shvec\textsuperscript{a}, Kostiantyn Pavlov\textsuperscript{b}, Olena Pavlova\textsuperscript{b}, Inna Irtyshcheva\textsuperscript{c}, Nazariy Popadynets\textsuperscript{d*}, Yevheniia Boiko\textsuperscript{e} and Iryna Kramarenko\textsuperscript{f}

\textsuperscript{a}National University of Water and Environmental Engineering, Ukraine
\textsuperscript{b}Lesya Ukrainka Eastern European National University, Ukraine
\textsuperscript{c}Admiral Makarov National University of Shipbuilding, Ukraine
\textsuperscript{d}SI "Institute of Regional Research named after M. I. Dolishniy of the NAS of Ukraine", Ukraine
\textsuperscript{e}Mykolaiv Interregional Institute of Human Development of the University "Ukraine", Ukraine

\textbf{ABSTRACT}

This paper deals with retrial queues of the $|\text{M}_q/\text{M}/c/\infty|$- type in which the rate of input flow depends on the number of sources of repeated calls and each call has only one retrial attempt. That is, if a call fails to enter the server facility at the retrial attempt, then it leaves the system without service. The existence conditions of stationary regime and the vector-matrix of the stationary probabilities of the service process are represented. This representation uses an approximation of the initial model by means of the truncated one and the directs passage to the limit. For these systems a threshold strategy for the rate of input flow is used. The multi-criterion optimization problem for finding optimal strategy of control is considered. The quality functionals of the optimization problem are represented through the stationary probabilities.

1. Introduction

Nowadays, the modern development of call centers, telecommunication and computer networks requires new models of queuing systems, which model real processes more fully. It is necessary to consider queuing systems that take into account the phenomenon of repeated calls. Such retrial queues can be used to solve the practical problems without loss of calculation accuracy. For retrial queues, it is assumed that calls that arrive in the system while all the servers are busy become sources of repeated calls and return their demands to get service after some time. Calls that had no access to servers at the moment of arrival are assumed to be in the orbit and independently apply for service in a random time (Falín et al., 1997; Kovalenko, Koba, 2010). All calls in the orbit are equivalent, disordered.

2. Literature review

Classical retrial queues have been analyzed in the monographs and articles where generating functions and truncated models have been used (Falín et al., 1997; Artalejo et al., 2008; Yakubiv et al., 2020; Kuznetsov et al., 2019; Pavlov., 2020). It made it possible to obtain a lot of results for such systems. Typically for retrial queues, it is assumed that each call can get its repeated demands until it receives service. This is only an approximation to real situations as the number of repeated attempts to the system is often limited. Systems with limited number of retrials attempts to begin service have been investigated numerically by using the algorithmic method and simulation experiments (Shin, Moon, 2008; Irtyshcheva et al., 2020; Harkusha et al., 2016; Hrynkevych et al., 2020). But research of these retrial queues also includes the search for conditions of the stationary mode existence and construction of algorithms for calculating the stationary distribution. It is
quite relevant in nowadays, in particular, from the point of view of optimization these systems. The problem of choice of the optimal rate of the input flow has its practical meaning. The threshold strategy is used for an optimal control in retrial queues. It can be used in the case of the rate of input flow dependent on the number of calls in the orbit (the number of sources of repeated calls). Such modifications of retrial queues are more difficult for research.

In the present paper, we consider the case of the rate of input flow dependent on the number of calls in the orbit. Also we assumed that each call can make a single retrial attempt to begin service process. Under these conditions, the method of generating functions is not used in the analysis of service process. To investigate the characteristics of service process in stationary mode, the approximate approach has been developed. According to this approach, a truncated finite system is considered first, and then the obtained results are used for analysis of the original system. The aim of this paper is to solve the problem of control of the rate of input flow in case of threshold strategy using effective formulas for stationary probabilities.

3. Materials and Methods

The main model that we consider in this paper is a two-dimensional Markov chain \( Q(t) = (Q_1(t), Q_2(t))^T \), \( t \geq 0 \) with continuous time in phase space \( S(Q) = \{0,1,\ldots,c\} \times \{0,1,\ldots\} \). The component \( Q_1(t) \) is the number of servers that are busy at the instant of time \( t \) and \( Q_2(t) \) is the number of sources of repeated calls. The infinitesimal characteristics \( a_{(i,j)}(i',j'),(i'',j'') \in S(Q) \) of chain \( Q(t) \) are specified as follows:

If \( i = 0,\ldots,c-1, j = 0,1,\ldots, \) then

\[
a_{(i,j)} = \begin{cases} \lambda_j, & \text{for} \quad (i',j') = (i+1,j), \\ i\mu, & \text{for} \quad (i',j') = (i-1,j), \\ jv, & \text{for} \quad (i',j') = (i+1,j-1), \\ -(\lambda_j + i\mu + jv), & \text{for} \quad (i',j') = (i,j), \\ 0, & \text{otherwise}; \end{cases}
\]

(1)

If \( i = 0,\ldots,c, j = 0,1,\ldots, \) then

\[
a_{(i,j)} = \begin{cases} \lambda_j, & \text{for} \quad (i',j') = (c,j+1), \\ c\mu, & \text{for} \quad (i',j') = (c-1,j), \\ jv, & \text{for} \quad (i',j') = (c,j-1), \\ -(\lambda_j + c\mu + jv), & \text{for} \quad (i',j') = (c,j), \\ 0, & \text{otherwise}. \end{cases}
\]

(2)

This Markov chain models the process of service in the following system. Calls arrive at the input for service. If there is at least one free server at the moment of arrival, the call instantly occupies it and then leaves the system after the service. The service time is a random variable exponentially distributed with the parameter \( \mu \). If all the servers are busy, then the arriving call becomes a source of repeated call and tries again to obtain service in a random time that is exponentially distributed with the parameter \( v \). The call that finds out all the servers busy at the time of retrial, leaves the system and does not obtain the service. It means that each call has only one retry to begin the service. The intensity of arrival flow is \( \lambda_j, j = 0,1,\ldots \) and depends on the number \( j \) of sources of repeated calls. Similar processes were analyzed by numerical methods in (Shin, Moon, 2008). According to the notation accepted in the queuing theory, we will represent the model under study by \( |Mg/M/c/\| \). The character \( \| \) at the last place means the absence of constraints for the number of sources of repeated calls.

We also assume that \( \lambda_j, \mu, v > 0, j = 0,1,\ldots, \) which is always taken into account in practice. If this requirement is satisfied, we will say that the model parameters are non-degenerate.

The conditions of the existence of stationary mode for \( Q(t) = (Q_1(t), Q_2(t))^T \) are non-degenerate model parameters and

\[
\lim_{n \to \infty} n^{-1} \lambda_n < v \quad \text{that were established in the lemma 1 (Pryshchepa, Lebedev, 2017)}.
\]

Denote by \( \pi_{ij}(i,j) \in S(Q) \), stationary distribution of the system. The stationary probabilities \( \pi_{ij}, i = 0,1,\ldots,c, j = 0,1,\ldots, \) satisfy the following Kolmogorov's system of equations

\[
(\lambda_j + jv)\pi_{ij} = \mu\pi_{i-1,j}, \quad j = 1,\ldots
\]

(3)

\[
(\lambda_j + i\mu + jv)\pi_{ij} = \lambda_j\pi_{i-1,j} + (i+1)\mu\pi_{i-1,j} + (j+1)v\pi_{i-1,j+1}, \quad i = 1,\ldots,c-1, \quad j = 0,1,\ldots
\]

(4)

\[
(\lambda_j + c\mu + jv)\pi_{ij} = \lambda_j\pi_{c-1,j} + \lambda_j\pi_{c,j-1} + (j+1)v\pi_{c-1,j+1} + (j+1)v\pi_{c,j+1}, \quad j = 0,1,\ldots
\]

(5)
and the normalizing condition
\[ \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \pi_{ij} = 1. \]

For each \( j=0,1,... \) construct partition of the phase space \( S(Q) = \bigcup_{j} \bigcup_{i} S_j(Q) \), where \( S_j(Q) = \{(k,m) \in S(Q) : m \leq j\} \).

Using the theorem about the equality of flows of probabilities through the boundary of domain \( S_j(Q) \) in stationary mode (Walrand, 1988), we get
\[ \lambda_j \pi_{ij} = (j+1) \nu \pi_{ij} + (j+1) \nu \pi_{i(j+1)} +...+(j+1) \nu \pi_{i(j+1)} \]

In the general case, for an arbitrary number of sources of repeated calls, it is impossible to find explicit formulas for the stationary probabilities that satisfy the Eqs. (3)-(7). An exception is one-channel systems, which are considered in (Lebedev, Pryshchepa, 2007). Taking this into account, let us consider a truncated model \(|MQ/M/c/N|\) which has a finite number of sources for repeated calls \( N \). If all the servers are busy and \( N \) places of repeated calls are generated, new arrivals are completely lost by the system. The process of service for truncated model is defined by two-dimensional Markov chain
\[ N 
\]
with continuous time in the phase space \( \times \). The infinitesimal characteristics \( \pi_{ij} \) of chain \( N \) coincide with (1) for \( i=0,1,...,c-1, j=0,1,...,N \) and coincide with (2) for \( i=c, j=0,1,...,N-1 \). If \( i=c, j=N \) we get
\[ \pi_{ij} \] and the normalizing condition
\[ \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \pi_{ij} = 1. \]

Let us introduce the notation:
\[ e_i = (1,0,...,0)_{i=1} \quad \tau_{i-1} = (1,1,...,1)_{i=1} \]

\( F(j) = \|f_{ik}\|_{k=0}^{\infty} \) is a matrix with the elements \( f_{ik} = \lambda_j + i \mu + j v, i = 0,...,c-1, f_{ik} = \lambda_j + c \mu, f_{ik} = -\lambda_j, \)
\( i = 0,...,c-1, f_{ik} = -(\lambda_j + j v), f_{ik} = -(i+1) \mu, f_{ik} = -j v, k = 0,1,...,c-2 \); the other elements are equal to zero;
\( B = \|b_{ik}\|_{k=0}^{\infty} \) is a matrix with the elements \( b_{ik} = 1, i = 1,2,...,c, b_{ck} = 1 \), the other elements being equal to zero; \( D = \|d_{ik}\|_{k=0}^{\infty} \) is the asymptotic representation of stationary probabilities for the model \(|MQ/M/c/\alpha|\).
Theorem 1 (Pryshchepa, Lebedev, 2017). If the model parameters are non-degenerate and \( \lim_{n \to \infty} n^{-1} \lambda_n < v \), then the stationary probabilities \( \pi_{ij}(i, j) \in S(\mathcal{Q}) \) of the system \(|M_0/M/c/\infty|\) can be found as follows:

\[
\pi_j = \Delta_i \pi_{i0}, \quad j = 0, 1, \ldots
\]

where

\[
\pi_{i0} = \left( \sum_{j=0}^{\infty} \Delta_j \right)^{-1} \quad \Delta_j = \lim_{n \to \infty} \frac{1}{j!v^j} \prod_{i=0}^{j-1} W(k) D^i e_{i+1} e_{i+1} \quad j = 0, 1, \ldots
\]

This theorem was proved by using the result for the vector of stationary probabilities \( \pi^{(N)}_{ij}(0, \ldots, T) \) that satisfy equations (8)-(14). It became possible because the service processes \( Q_t \) and \( Q_N(t) \) are migration processes. Therefore, on the basis of the results from [1, Sec. 2, Theorems 2.3 and 2.4] as \( N \to \infty \), the stationary probabilities \( \pi^{(N)}_j \) of the truncated model \(|M_0/M/c/N|\) approximate the stationary probabilities \( \pi_j \) of the model \(|M_0/M/c/\infty|\).

Optimal control in a set of threshold strategies.

We have analyzed retrial queues with a single retrial attempt in case of the rate of input flow dependent on the number of calls in the orbit. It makes it possible to set and solve optimization problems for such systems. In view of this we consider a class of many-threshold strategies that are set by thresholds \( 0 = h_0 < h_1 < \ldots < h_K = \infty \), \( h = (h_1, \ldots, h_K) \), \( K \) is a fixed number. If at the \( t \geq 0 \) moment of time the number of retrial calls is in interval \((H_{i-1}, H_i] \), \( i = 1, \ldots, K \), then the system operates in the \( i \)th regime and the input flow rate is equal to \( h_i \). If the number of retrial calls is more than \( H_i \), then the system operates in the next regime with the input flow rate \( h_{i+1} \). Other parameters do not depend on the regime. The use of the threshold strategy allows us to solve a set of economical problems. The high rate of the input flow leads to an increase of sources of repeated calls. Through this, calls can leave the system without its service. Thus, the input flow rate should decrease. In the case of a small rate of the input flow, the number of calls in the service area will be reduced and the servers will be free. This, in its turn, will affect the economic indicators of the systems. In this case, it is necessary to increase the rate of the input flow. Formally, the threshold strategy \( h \) fixes the dependence \( \lambda_j \) on the number of retrial calls.

\[
\lambda_j = \begin{cases} 
1, & j = 0, \ldots, H_1; \\
h_2, & j = H_1 + 1, \ldots, H_2; \\
\vdots & \vdots \\
h_K, & j = H_{K-1} + 1, \ldots.
\end{cases}
\]

The optimal control problem for the retrial queues with a single retrial attempt consists in searching such thresholds \( H_i \), \( i = 1, \ldots, K-1 \), that are solutions of multi-criterion optimization problem. Such optimization problem requires at least two functionals of quality. For example, the first is to maximize profits and the second is to minimize costs. We can consider multi-criterion optimization problems that contain more functionals of quality.

So we consider the statement of the multi-criterion optimization problem. Let us denote \( S_i(t, H) \) is the number of calls served to \( t \) moment in the \( i \)th regime, \( i = 1, \ldots, K, \)

\( \tilde{S}_K(t, H) \) is the number of calls which were rejected and transformed in retrial calls;

\( \bar{S}_K(t, H) \) is the number of switchings of the rate of the input flow;

\( \tilde{S}_K(t, H) \) is the number of calls that leave the system without service after repeated attempt.

The limits \( \lim_{t \to \infty} t^{-1} S_i(t, H) \) exist for the queue with non-degenerate parameters. We denote them via \( S_i(H), i = 1, \ldots, K+3 \).

Formally, optimization problem can be represented as follows:

\[
S_i(H) \to \max, \quad i = 1, 2, \ldots, K, \\
\tilde{S}_i(H) \to \min, \\
\bar{S}_i(H) \to \min, \\
\tilde{S}_i(H) \to \min, \\
H_i \leq H_{i+1}, \quad H_i \in \{0, 1, \ldots\}, \quad i = 1, 2, \ldots, K-1.
\]

As the main approach to the solution of the multi-criterion problem (15), we use the method of linear convolution. This method leads to the following optimization problem:
where the cost coefficients are: $C_i$ is the income relating to the service of a call in the $i$-th regime, $i=1,...,K$; $C_{K+1}$ is the penalty relating to service failure; $C_{K+2}$ is the penalty relating to switching of the input flow rate, $C_{K+3}$ is the penalty relating to lost call after it’s the last failed attempt. The solution of the problem (16) is such many-threshold strategy that maximizes the average system operation profit. The functionals $S_i(H), i=1,...,K+3,$ can be written via the stationary probabilities of service process:

\[
\begin{align*}
L_i(H) &= \mu \sum_{j=H_i+1}^{\infty} n \pi_{ij}(H) \\
L_{K}(H) &= \mu \sum_{j=1}^{H_{K+1}-1} \sum_{k=H_{K+1}}^{\infty} n \pi_{ik}(H) \\
L_{K+1}(H) &= h_1 \sum_{j=1}^{H_{K+1}} \pi_{ij}(H) + h_2 \sum_{j=H_{K+1}+1}^{\infty} \pi_{ij}(H) + \ldots + h_K \sum_{j=H_{K+1}}^{\infty} \pi_{ij}(H) \\
L_{K+2}(H) &= \sum_{j=1}^{K} h_j \pi_{ij}(H) + (H_{K+1}+1) \sum_{n=0}^{\infty} \pi_{nH_{K+1}}(H) \\
L_{K+3}(H) &= \nu \sum_{j=1}^{K} j \pi_{ij}(H)
\end{align*}
\]

In the case $K=2$, formulation of the optimization problem becomes simpler as follows,

\[
S(H) = C_1 S_1(H) + C_2 S_2(H) - C_4 S_4(H) - C_5 S_5(H) \rightarrow \max \quad H \in \{0, 1, \ldots\}
\]

Thus, the functionals $S_i(H), i=1,5$ can be represented in terms of the stationary probabilities of $|M_{\infty}/M/c/\infty|$-queue as follows:

\[
\begin{align*}
L_1(H) &= \mu \sum_{j=H+1}^{\infty} n \pi_{ij}(H) \\
L_2(H) &= \mu \sum_{j=1}^{H_{2+1}-1} \sum_{k=H_{2+1}}^{\infty} n \pi_{ik}(H) \\
L_3(H) &= \mu \sum_{j=0}^{H_{3+1}} \sum_{k=H_{3+1}}^{\infty} n \pi_{ik}(H) \\
L_4(H) &= h_1 \pi_{iH}(H) + (H_{2+1}+1) \sum_{n=0}^{\infty} \pi_{nH_{2+1}}(H) \\
L_5(H) &= \nu \sum_{j=1}^{K} j \pi_{ij}(H)
\end{align*}
\]

4. Results and discussion

As an example, we present the optimization problem for the $|M_{\infty}/M/c/\infty|$- system with the following parameters: $K=2$, $h_1=8$, $h_3=3$, $\mu=5, v=1$. Let $C_1=100$, $C_2=75$, $C_3=5$, $C_5=42$, are values of the cost coefficients. The maximum value of the quality functional is $S(H) = 156.69$ and it is obtained at the value of threshold $H=3$ (See Fig. 1). Also in the case of the threshold control strategy of input flow rate we present the $|M_{\infty}/M/c/\infty|$- system with the following parameters: $K=2$, $h_1=8$, $h_3=3$, $\mu=3, v=3$. Let $C_1=96$, $C_2=75$, $C_3=8$, $C_5=42$. The maximum value of the quality functional is $S(H) = 251.42$ and it is obtained at the value of threshold $H=4$ (Fig. 2).

Fig. 1. Dependency curve of the quality functional on the control strategy for the $|M_{\infty}/M/1/\infty|$ system

Fig. 2. Dependency curve of the quality functional on the control strategy for the $|M_{\infty}/M/2/\infty|$ system

Source: calculated by the authors.

5. Conclusion

In the present paper, we have investigated retrial queues with one retry to begin service. We have represented formulas for stationary probabilities in terms of systems’ parameters, which allow solving optimization problems. In accordance with threshold strategy, the multi-criterion optimization problem of the effective functioning of the system has been formulated and solved. Quality functionals of the optimization problem are represented through the stationary probabilities.
References


© 2020 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).