

Steady state analysis of the M/G/1//N queue with orbit of blocked customers

Velika I. Dragieva¹

Published online: 30 September 2015
© Springer Science+Business Media New York 2015

Abstract This article considers a finite-source queueing model of M/G/1 type in which a customer, arriving at a moment of a busy server, is not allowed either to queue or to do repetitions. Instead, for an exponentially distributed time interval he is blocked in the orbit of inactive customers. We carry out a steady state analysis of the system and compare it with the corresponding system with retrials. Optimization problems are considered and formulas for the Laplace–Stieltjes transform of the busy period length are obtained.

Keywords Finite queues · Blocked customers · Losses · Retrials

1 Introduction

In queueing theory the source/sources of demands (calls) usually remain hidden behind the term input flow. In many practical situations it is appropriate to assume that the input flow properties do not change during the system functioning and do not depend on the number of demands under service or waiting for service. A good approximation of such situations is given by the queueing models with infinite source, especially the models with Poisson input flow. However, there are situations in which the assumption of an unchanging input flow is unacceptable. Basically, these are the cases when the server/servers serve a finite number of customers, each one producing his/her own flow of demands. In these systems the generalized input flow depends on the number of customers able to produce demands, i.e. the customers not being under service or not waiting for service. In order to approximate such systems we need finite-source (also called quasi-random input or closed) queueing models. These models arise in various practical areas as local area networks (Janssens 1997; Li and Yang 1995; Mehmet-Ali et al. 1988), cellular mobile networks (Artalejo and Gómez-Corral 2007; Tran-Gia and Mandjes 1997; Van Do et al. 2014), magnetic disk memory systems

✉ Velika I. Dragieva
vildrag2001@yahoo.com

¹ University of Forestry, 1756 Sofia, Bulgaria

(Ohmura and Takahashi 1985), bufferless optical networks (Overby 2005; Zukerman et al. 2004), mobile networks that provide packet radio service (Dahmouni et al. 2005).

A customer, unable to get service at a moment of his/her arrival (because of a busy server/servers, server vacation or repair, etc.) can behave in different ways. In many real situations such customers repeat their attempts in pre-determined or random intervals until they receive service. The appropriate queueing models for such situations are the models with retrials (repetitions, returning customers). In the past twenty years there is a large number of articles devoted to retrial queues. The reader can find a detailed review of the main results, methods of analysis and literature on retrial queues in the monographs by Falin and Templeton (1997) and Artalejo and Gómez-Corral (2007).

We can find queues with returning customers in our daily activities as well as in many telephone and communication systems. Most of the models with a finite number of sources described above are also models with retrials. The simplest example of such model is the situation when a telephone subscriber gets a busy signal and repeats the call until the demand is satisfied. These repeated attempts form an additional flow, along with the existing stream of primary calls. As a result, the total flow of calls circulating in the telephone network consists of two flows: one, which reflects the real wishes of the telephone subscribers and the other, which arises as a consequence of the lack of success of previous attempts. To analyze this changed total flow and the corresponding queueing system we have to use models with retrials.

Let us now consider the same example, but under the assumption that the system operator may temporarily prohibit access to the system for all unsuccessful subscribers. In other words, in some situations (high intensity of the input flow or of the server utilization, etc.), all subscribers which obtain a busy signal are “blocked” for a pre-determined or random time interval, during which they are not allowed to make new calls, either repeated, or primary. Will this in fact improve the system performance? And in which cases exactly?

This problem is a particular case of the optimization problems (like optimal control, cost functions, etc.), which are one of the main motivators and tools for constructing and analyzing competitive queueing models (see for example Artalejo and Phung-Duc 2012; Efrrosinin and Breuer 2006; Kim et al. 2006 and the references therein). This problem in fact is one of the main motivations for the analysis presented in this paper. We consider a single-server queueing system with a finite number of customers (sources of demands). The server can be in two states: free (idle) and busy (working). Each one of the customers produces a Poisson flow of calls of the same intensity. If the server is busy at the time moment of a call arrival, the customer is blocked for an exponentially distributed time interval, during which he/she can't produce demands. All blocked customers are said to be in the orbit of blocked or inactive customers, in the inactive orbit. After the blocking interval is over, the customer is free to produce new calls.

This system can be considered as a variant of the model with retrials or of the model with losses. Its investigation is interesting both in itself and as a tool for optimization of the corresponding retrial model. In this model, i.e. a single server finite retrial queue, the failed customers, instead of being blocked, repeat their attempts in exponentially distributed time intervals. The model is useful for performance analysis of many real queueing systems and has been extensively studied in a number of papers (Amador 2010; De Kok 1984; Dragieva 2013; Falin and Artalejo 1998; Ohmura and Takahashi 1985).

To the best of our knowledge, most of the studies on finite queues with losses consider Engset models and are mainly concerned with formulas for the blocking probability in the multi-server queues with exponential service times (Dahmouni et al. 2005; Moscholios and Logothetis 2006; Overby 2005; Zukerman et al. 2004). In these models it is assumed that

the customers are ready to produce new calls immediately after the failures. In many real situations this is not realistic. The customers are lost from the system management point of view, but they need to satisfy their demands in some way, for example by another firm, or operator, etc. They will hardly have new requests before the completion of the previous one. During this time the customers should be considered as inactive or missing from the system. In finite queues these missing customers change the input flow of demands and have to be taken into account, as it is in our model with inactive orbit.

The rest of the paper is organized as follows. Section 2 describes the considered model in details. Section 3 contains the main results of the paper: formulas for computing the stationary distributions of the system states and the basic performance macro characteristics. In Sect. 4 we consider some asymptotic properties of these characteristics and in Sect. 5 we give numerical examples. Optimization problems are discussed in Sect. 6. In Sect. 7, applying the discrete transformations, obtained in Sect. 3, we derive formulas for computing the Laplace-Stieltjes transform of the distribution of the busy period length. In fact, the results, presented in this section refer to the analysis of the system in non-stationary regime. Section 8 concludes the paper.

2 Model description

The queueing model under consideration has one server which serves N , $2 \leq N < \infty$ customers, also called sources of demands (requests, calls). Each of these customers can be in one of the following three states: active (or free), under service, and inactive (or blocked).

If the customer is active, it produces a Poisson flow of calls with rate λ . This means that when a source is active at time moment t it may generate a call during time interval $(t, t + dt)$ with probability λdt . Thus, if at instant t there are n active customers, then the probability that a call arrives during interval $(t, t + dt)$ is equal to $n\lambda dt$.

If the server is free at the time of a call arrival, then the call is immediately served and once the service is over, the source becomes active again. During the service time the source cannot generate a new call.

If the server is busy at the instant of a call arrival, the source moves into inactive state and stays in this state for a random time interval, exponentially distributed with parameter μ . While being inactive, the source cannot generate a call. When the time of inactivity is over, the source moves again into the active state and is free to generate a call. All inactive customers are said to be in the inactive orbit or to be blocked in the orbit. Thus, if at time moment t there are n customers in the orbit, then the probability that during a time interval $(t, t + dt)$ one of them moves into active state, is equal to $n\mu dt$.

The parameters λ and μ will be called *source arrival* and *source activation rates*, respectively.

The service times have probability distribution function $G(x)$, with $G(0) = 0$, hazard rate function

$$\gamma(x) = \frac{G'(x)}{1 - G(x)},$$

Laplace-Stieltjes transform— $g(s)$ and first moment— v^{-1} .

The input flow of calls, times of inactivity and service times are assumed to be mutually independent.

We analyze the system with the help of the supplementary variables method, and according to this method we describe the system states by means of the Markov process

$$\{X(t), t \geq 0\} = \{(C(t), R(t), z(t)), t \geq 0\}.$$

Here $C(t)$ is the number of busy servers at instant t (i.e. $C(t)$ is 0 or 1 according to whether the server is free or busy at time moment t), $R(t)$ is the number of inactive customers at instant t (orbit size), $z(t)$ is the supplementary variable introduced in the case $C(t) = 1$ and equal to the elapsed service time. From the model description it is clear that the situation $R(t) = N$ is impossible both for $C(t) = 1$ and for $C(t) = 0$. Thus, the possible values of the orbit size, $R(t)$ are $\{0, 1, \dots, N - 1\}$.

We define the probabilities (densities)

$$p_{1j}(t, x)dx = P \{C(t) = 1, R(t) = j, x \leq z(t) < x + dx\}, \quad (1)$$

$$p_{ij}(t) = P \{C(t) = i, R(t) = j\}, \quad i = 0, 1, \quad j = 0, 1, \dots, N - 1 \quad (2)$$

and, following the method of supplementary variables derive the equations of statistical equilibrium for the limit probabilities as $t \rightarrow \infty$:

$$\begin{aligned} \frac{d}{dx} p_{1j}(x) = & -[(N - j - 1)\lambda + j\mu + \gamma(x)]p_{1j}(x) \\ & + (j + 1)\mu p_{1,j+1}(x) + (N - j)\lambda p_{1,j-1}(x), \end{aligned} \quad (3)$$

$$[j\mu + (N - j)\lambda]p_{0j} = (j + 1)\mu p_{0,j+1} + \int_0^\infty p_{1j}(x)\gamma(x)dx, \quad (4)$$

$$p_{1j}(0) = (N - j)\lambda p_{0j}, \quad j = 0, 1, \dots, N - 1, \quad (5)$$

$$p_{1,-1}(x) = p_{1N}(x) = p_{0N} = 0,$$

$$\sum_{j=0}^{N-1} p_{0j} + \sum_{j=0}^{N-1} \int_0^\infty p_{1j}(x)dx = 1. \quad (6)$$

It should be noted that because of the finite state space of the process $(C(t), R(t))$ the limit probabilities $p_{1j}(x)$, p_{1j} , p_{0j} that satisfy Eqs. (3)–(6) always exist. Formulas for their calculation are derived in the next section.

3 Stationary distributions of the system states

In this section we derive formulas for computing the stationary joint distributions p_{in} of the server state and the orbit size and the main macro characteristics of the system performance. To this end, we first apply the method of discrete transformations and find the solutions of Eq. (3). It should be noted that this method is common in the analysis of finite queueing models (Falín and Artalejo 1998; Jaiswal 1969; Ohmura and Takahashi 1985; Wang et al. 2011; Zhang and Wang 2013), and can be considered as a particular case of the eigenvalue method, applicable in the analysis of various queueing models (see for example Drekić and Grassmann 2002; Lee et al. 2005 and the references therein).

We now rewrite Eq. (3) in a matrix form

$$[\theta I - A] \bar{p}_1(x) = 0, \quad (7)$$

where

$$\theta = \gamma(x) + \frac{d}{dx},$$

I is the identity matrix of order N , A is constructed from (3) in the usual way and $\bar{p}_1(x)$ is the column vector of the unknown functions $p_{1j}(x)$,

$$\bar{p}_1(x) = (p_{10}(x), \dots, p_{1,N-1}(x))^T.$$

Then, in the next Proposition 1 we derive formulas for computing the entries of the matrices Y and Λ , such that

$$Y^{-1}AY = \Lambda. \tag{8}$$

Proposition 1 *The matrix Λ is a diagonal one, $\Lambda = \text{diag}\{\lambda_0, \lambda_1, \dots, \lambda_{N-1}\}$ with*

$$\lambda_i = -i(\mu + \lambda), \quad i = 0, 1, \dots, N - 1$$

and the entries of the k th column of Y , $y^{(k)} = (y_0^{(k)}, \dots, y_{N-1}^{(k)})^T$, $k = 0, 1, \dots, N - 1$, can be calculated by the relations

$$y_0^{(k)} = 1, \tag{9}$$

$$y_n^{(k)} = \frac{-k(\lambda + \mu)}{n\mu} (y_0^{(k)} + \dots + y_{n-1}^{(k)}) + \frac{(N - n)\lambda}{n\mu} y_{n-1}^{(k)}, \tag{10}$$

$$n = 1, \dots, N - 1,$$

or by their equivalent formulas

$$y_n^{(k)} = \sum_{i=0}^n (-1)^{n-i} \left(\frac{\lambda}{\mu}\right)^i \binom{N - k - 1}{i} \binom{k}{n - i}, \tag{11}$$

with

$$\binom{j}{l} = 0 \text{ if } l > j.$$

Furthermore, for the sum of the first n coordinates of the k th column we have

$$\sum_{i=0}^n y_i^{(k)} = \begin{cases} \sum_{i=0}^n \left(\frac{\lambda}{\mu}\right)^i \binom{N-1}{i} & \text{for } k = 0 \\ \sum_{i=0}^n (-1)^{n-i} \left(\frac{\lambda}{\mu}\right)^i \binom{N-k-1}{i} \binom{k-1}{n-i} & \text{for } k = 1, \dots, N - 1 \end{cases} \tag{12}$$

and therefore

$$\sum_{i=0}^{N-1} y_i^{(k)} = \begin{cases} \left(1 + \frac{\lambda}{\mu}\right)^{N-1} & \text{for } k = 0 \\ 0 & \text{for } k = 1, \dots, N - 1 \end{cases} \tag{13}$$

Proof The matrices Λ and Y (Eq. (8)) depend on the eigenvalues and the eigenvectors of A . For the matrix $A - tI$ we have

$$A - tI = \begin{pmatrix} (1 - N)\lambda - t & \mu & 0 & \dots & 0 & 0 \\ (N - 1)\lambda & (2 - N)\lambda - \mu - t & 2\mu & \dots & 0 & 0 \\ 0 & (N - 2)\lambda & (3 - N)\lambda - 2\mu - t & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda & -(N - 1)\mu - t \end{pmatrix}$$

Now we add the first row of this matrix to the second one, then—the newly formed second row—to the third one, and so on. As a result the matrix $A - tI$ takes the equivalent form

$$A - tI \sim \begin{pmatrix} (1 - N)\lambda - t & \mu & 0 & \dots & 0 & 0 & 0 \\ -t & (2 - N)\lambda - t & 2\mu & \dots & 0 & 0 & 0 \\ -t & -t & (3 - N)\lambda - t & \dots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -t & -t & -t & \dots & -t & -\lambda - t & (N - 1)\mu \\ -t & -t & -t & \dots & -t & -t & -t \end{pmatrix}$$

From this expression it follows that $t = 0$ is an eigenvalue of A and that one of the corresponding eigenvectors is the vector $y^{(0)} = (y_0^{(0)}, \dots, y_{N-1}^{(0)})^T$ with coordinates

$$y_0^{(0)} = 1, \dots, \\ y_n^{(0)} = \frac{(N - n)\lambda y_{n-1}^{(0)}}{n\mu} = \binom{N - 1}{n} \left(\frac{\lambda}{\mu}\right)^n, \quad n = 1, \dots, N - 1.$$

This proves formulas (9), (10) for $k = 0$.

Further, in the equations $[A - t_k I] y^{(k)} = 0$ we express the n th coordinate of $y^{(k)}$ in terms of the previous ones,

$$y_1^{(k)} = \frac{(N - 1)\lambda y_0^{(k)} + t_k y_0^{(k)}}{\mu}, \dots, \\ y_n^{(k)} = \frac{t_k [y_0^{(k)} + y_1^{(k)} + \dots + y_{n-1}^{(k)}] + (N - n)\lambda y_{n-1}^{(k)}}{n\mu}, \\ n = 2, \dots, N - 1.$$

If in these relations we choose $y_0^{(k)} = 1$ and substitute $t_k = -k(\lambda + \mu)$, we obtain (9), (10) for any $k, 1 \leq k \leq N - 1$. Then, using some combinatorial formulas we derive (11)–(13). Equation (13) shows that $\det(A - t_k I) = 0$ for each $t_k = -k(\lambda + \mu)$. This finishes the proof of Proposition 1.

Now, having the matrices A and Y we consider the transformations

$$\bar{p}_1(x) = Y\bar{q}_1(x) \tag{14}$$

and applying them in the matrix equation (7) we get it in a simpler form

$$[\theta I - \Lambda]\bar{q}_1(x) = 0. \tag{15}$$

With the help of (14) and (15), on the basis of (3)–(6), we derive formulas for the stationary system state distributions $p_{1n}(x), p_{0n}, p_{1n}$. They are given in the next theorem.

Theorem 1 *The stationary joint distributions $p_{1n}(x), p_{0n}, p_{1n}, n = 0, \dots, N - 1$ of the server state and the orbit size in the considered $M/G/1//N$ queue with inactive orbit can be calculated by the formulas*

$$p_{1n}(x) = \sum_{k=0}^{N-1} y_n^{(k)} q_{1k}(0)(1 - G(x))e^{-k(\lambda+\mu)x}, \tag{16}$$

$$p_{1n} = y_n^{(0)} \frac{q_{10}(0)}{\nu} + \sum_{k=1}^{N-1} y_n^{(k)} q_{1k}(0) \frac{1 - g_k}{k(\lambda + \mu)}, \tag{17}$$

$$p_{0n} = \frac{1}{(N - n)\lambda} \sum_{k=0}^{N-1} y_n^{(k)} q_{1k}(0). \tag{18}$$

Here:

- $y_n^{(k)}$ are given in Proposition 1, ((9)-(12));
- $g(s)$ is the Laplace-Stieltjes transformation of the service distribution function, $G(x)$ and

$$g_k = g(k(\lambda + \mu));$$

- the quantities $q_{1k}(0)$ are solutions of the following system of $(N - 1)$ linear equations:

$$\sum_{k=0}^{N-1} \left[(1 - g_k) \sum_{i=0}^n y_i^{(k)} - \frac{(n + 1)\mu}{(N - n - 1)\lambda} y_{n+1}^{(k)} \right] q_{1k}(0) = 0, \tag{19}$$

$n = 0, \dots, N - 2$ with a normalizing condition

$$\begin{aligned} &\sum_{n=0}^{N-1} y_n^{(0)} \left[\frac{1}{(N - n)\lambda} + \frac{1}{\nu} \right] q_{10}(0) \\ &+ \sum_{k=1}^{N-1} \sum_{n=0}^{N-1} y_n^{(k)} \left[\frac{1}{(N - n)\lambda} + \frac{1 - g_k}{k(\lambda + \mu)} \right] q_{1k}(0) = 1. \end{aligned} \tag{20}$$

Proof As written above, we apply transformations (14) in the Eq. (7) and obtain (15). Thus, we have a system of ordinary differential equations,

$$\frac{d}{dx} q_{1n}(x) = - [\gamma(x) + n(\lambda + \mu)] q_{1n}(x),$$

with solutions

$$q_{1n}(x) = q_{1n}(0) (1 - G(x)) e^{-n(\lambda + \mu)x}, \quad n = 0, \dots, N - 1.$$

From the last formula and relations (14) we obtain (16). Formulas (17) follow from (16) and relations

$$p_{1n} = \int_0^\infty p_{1n}(x) dx.$$

From Eq. (5) we have

$$p_{0n} = \frac{p_{1n}(0)}{(N - n)\lambda} \tag{21}$$

and substituting here $p_{1n}(0)$ according to (16) we get (18). Now, from (4) and (6), substituting p_{0n} according to (21) we have

$$\begin{aligned} &\left[1 + \frac{n\mu}{(N - n)\lambda} \right] p_{1n}(0) = (1 - \delta_{n,N-1}) \frac{(n + 1)\mu}{(N - n - 1)\lambda} p_{1,n+1}(0) \\ &+ \int_0^\infty p_{1n}(x) \gamma(x) dx, \\ &\sum_{n=0}^{N-1} \left[\frac{p_{1n}(0)}{(N - n)\lambda} + \int_0^\infty p_{1n}(x) dx \right] = 1. \end{aligned}$$

Here δ_{ij} is Kronecker’s delta, equal to 1, if $i = j$, and equal to 0, if $i \neq j$. In each of the last N equations we substitute $p_{1n}(x)$ and $p_{1n}(0)$ according to (16) and obtain

$$\begin{aligned} & \sum_{k=0}^{N-1} q_{1k}(0) \left[y_n^{(k)} \left(\frac{n\mu}{(N-n)\lambda} + 1 - g_k \right) \right. \\ & \quad \left. - (1 - \delta_{n,N-1}) y_{n+1}^{(k)} \frac{(n+1)\mu}{(N-n-1)\lambda} \right] = 0, \tag{22} \\ & \sum_{n=0}^{N-1} y_n^{(0)} \left[\frac{1}{(N-n)\lambda} + \frac{1}{\nu} \right] q_{10}(0) \\ & \quad + \sum_{k=1}^{N-1} \sum_{n=0}^{N-1} y_n^{(k)} \left[\frac{1}{(N-n)\lambda} + \frac{1-g_k}{k(\lambda+\mu)} \right] q_{1k}(0) = 1. \end{aligned}$$

The last equation is just the normalizing condition (20), and summing up the first n of Eq. (22), $n = 1, \dots, N - 2$, we get relations (19). For $n = N - 1$ this summing leads to identity. This completes the proof of the theorem.

Thus, to calculate the stationary system state distributions we need the solutions $q_{1k}(0)$, $k = 0, 1, \dots, N - 1$ of the system (19)–(20). They can be easily obtained. For example, we first express from (19) all $q_{1k}(0)$, $k = 0, 1, \dots, N - 2$ in terms of the last one— $q_{1,N-1}(0)$ and, substituting in (20) we find the value of $q_{1,N-1}(0)$. Then, using relations (16)–(18) we can calculate any of the probabilities (densities) $p_{1n}(x)$, p_{0n} , p_{1n} and the basic macro characteristics of the system:

1. The server utilization: $P_1 = p_{10} + p_{11} + \dots + p_{1,N-1}$.
2. The probability that the server is idle: $P_0 = p_{00} + p_{01} + \dots + p_{0,N-1} = 1 - P_1$.
3. The mean number of inactive customers (mean orbit size):

$$E[R] = \sum_{n=1}^{N-1} n p_n = \sum_{n=1}^{N-1} n (p_{0n} + p_{1n}).$$

4. The mean rate of call generation:

$$\bar{\lambda} = \lambda E[N - C(t) - R(t)] = \lambda \{N - P_1 - E[R]\}.$$

We can also calculate the blocking probability that an arriving source finds the server busy and is blocked in the orbit of inactive customers. To this end we introduce conditional probabilities

$$\pi_{in} = \lim_{t \rightarrow \infty} P\{R(t) = n, C(t) = i | A(t)\},$$

where $A(t)$ is the event that at time moment t a source arrives. It is easy to verify that

$$\pi_{in} = \frac{(N - i - n) p_{in}}{N - P_1 - E[R]}, \quad i = 0, 1, \quad n = 0, \dots, N - 1$$

and that the blocking probability, P_B , is equal to

$$P_B = \sum_{n=0}^{N-1} \pi_{1n} = \frac{(N - 1)P_1 - \sum_{n=1}^{N-1} n p_{1n}}{N - P_1 - E[R]}.$$

4 Asymptotic properties as $\mu \rightarrow \infty$

In this section we derive the limit distributions of the system states and the limit values of the main performance characteristics as $\mu \rightarrow \infty$. If we denote

$$\tilde{y}_n^{(k)} = \lim_{\mu \rightarrow \infty} y_n^{(k)}, \tilde{p}_{in} = \lim_{\mu \rightarrow \infty} p_{in}, \tilde{q}_{1k}(0) = \lim_{\mu \rightarrow \infty} q_{1k}(0),$$

then it is easy to prove the following Proposition.

Proposition 2 *The limit probabilities \tilde{p}_{in} as $\mu \rightarrow \infty$ are equal to:*

$$\tilde{p}_{00} = \frac{\nu}{\nu + N\lambda}, \tilde{p}_{10} = \frac{N\lambda}{\nu + N\lambda}, \tag{23}$$

$$\tilde{p}_{0n} = \tilde{p}_{1n} = 0, n = 1, \dots, N - 1. \tag{24}$$

Proof Taking limits as $\mu \rightarrow \infty$, from Eqs. (9), (10) and (17)–(20) we obtain formulas for computing the corresponding limit values:

$$\tilde{y}_0^{(k)} = 1, \tilde{y}_n^{(k)} = \frac{-k}{n} \left(\tilde{y}_0^{(k)} + \dots + \tilde{y}_{n-1}^{(k)} \right),$$

and consequently

$$\tilde{y}_n^{(k)} = \begin{cases} (-1)^n \binom{k}{n} & \text{for } 0 \leq n \leq k \\ 0 & \text{for } k < n \leq N - 1 \end{cases} \tag{25}$$

$$k = 0, 1, \dots, N - 1, \tilde{p}_{1n} = \tilde{y}_n^{(0)} \frac{\tilde{q}_{10}(0)}{\nu} = \begin{cases} \frac{\tilde{q}_{10}(0)}{\nu} & \text{for } n = 0 \\ 0 & \text{for } 1 \leq n \leq N - 1 \end{cases} \tag{26}$$

$$\tilde{p}_{0n} = \frac{1}{(N - n)\lambda} \sum_{k=0}^{N-1} \tilde{y}_n^{(k)} \tilde{q}_{1k}(0) = \frac{1}{(N - n)\lambda} \sum_{k=n}^{N-1} \tilde{y}_n^{(k)} \tilde{q}_{1k}(0). \tag{27}$$

Here the quantities $\tilde{q}_{1k}(0)$ solve the following system of linear equations

$$\tilde{y}_{n+1}^{(0)} \tilde{q}_{10}(0) + \tilde{y}_{n+1}^{(1)} \tilde{q}_{11}(0) + \dots + \tilde{y}_{n+1}^{(N-1)} \tilde{q}_{1,N-1}(0) = 0, \tag{28}$$

$n = 0, \dots, N - 2$ with a normalizing condition

$$\tilde{q}_{10}(0) \sum_{n=0}^{N-1} \tilde{y}_n^{(0)} \left[\frac{1}{(N - n)\lambda} + \frac{1}{\nu} \right] + \sum_{k=1}^{N-1} \sum_{n=0}^{N-1} \frac{\tilde{y}_n^{(k)}}{(N - n)\lambda} \tilde{q}_{1k}(0) = 1. \tag{29}$$

Using (25), from (28) we obtain that all of the quantities $\tilde{q}_{1k}(0)$, with the exception of the first one, $\tilde{q}_{10}(0)$, are equal to zero

$$\tilde{q}_{11}(0) = \tilde{q}_{12}(0) = \dots = \tilde{q}_{1,N-1}(0) = 0.$$

Then, substituting in the normalizing condition (29) we obtain an equation for this nonzero quantity, $\tilde{q}_{10}(0)$,

$$\tilde{q}_{10}(0) \sum_{n=0}^{N-1} \tilde{y}_n^{(0)} \left[\frac{1}{(N - n)\lambda} + \frac{1}{\nu} \right] = 1.$$

Since $\tilde{y}_n^{(0)} \neq 0$ only when $n = 0$, we get

$$\tilde{q}_{10}(0) = \frac{N\lambda\nu}{\nu + N\lambda}.$$

Substituting in (26) and (27) we get (23), (24), which completes the proof of Proposition 2.

With the help of Proposition 2 we obtain formulas for the limit values:

$$\begin{aligned} \tilde{P}_0 &= \lim_{\mu \rightarrow \infty} P_0 = \frac{\nu}{\nu + N\lambda}, \quad \tilde{P}_1 = \lim_{\mu \rightarrow \infty} P_1 = \frac{N\lambda}{\nu + N\lambda}, \\ \lim_{\mu \rightarrow \infty} E[R] &= 0, \\ \tilde{P}_B &= \lim_{\mu \rightarrow \infty} P_B = \frac{(N-1)\tilde{P}_1 - \sum_{n=1}^{N-1} n\tilde{p}_{1n}}{N - \tilde{P}_1} = \frac{(N-1)\lambda}{(N-1)\lambda + \nu}, \\ \tilde{\lambda} &= \lim_{\mu \rightarrow \infty} \bar{\lambda} = \lambda E[N - C(t) - R(t)] = \lambda \left(N - \frac{N\lambda}{\nu + N\lambda} \right). \end{aligned}$$

5 Numerical examples

In this section we present numerical examples to illustrate graphically the influence of the system parameters on the performance characteristics, considered in the previous sections.

Figure 1 shows the influence of the source arrival rate, λ and the source activation rate, μ on the distribution p_{0n} of the orbit size when the server is idle, $n = 0, \dots, N - 1$. The presented results are calculated for four different distributions of the service time with the same mean, $1/\nu$:

- Deterministic distribution, equal to $1/\nu$, presented with dashed lines;
- Erlang distribution with parameters 4 and 4ν , presented with lines of stars;
- Exponential distribution with parameter ν , presented with solid lines;
- Uniform distribution in the interval $(0, 2/\nu)$, presented with lines of triangles. We can see that the distribution p_{0n} has only one mode and that when λ increases, the mode approaches $N - 1$, while when μ increases it is close to 0. This observation agrees with Proposition 2. We did not present here examples for p_{1n} and $p_n = p_{0n} + p_{1n}$, but they show that these distributions possess the same properties.

Figure 2 depicts the behaviour of the probability that the server is idle, P_0 versus each one of the system parameters:

- Source arrival rate, λ (the upper-left corner);
- Source activation rate, μ (the upper-right corner);
- Mean service time, $1/\nu$ (the lower-left corner);
- Number of customers, N (the lower-right corner).

The results presented here are calculated for the same four distributions of the service time, and are depicted with the same types of lines, as in Fig. 1.

In addition, the lines of dots show the values of P_0 in the corresponding finite queue with retrial orbit. They are calculated for the exponential distribution of the service time, according to the formulas, given in Falin and Artalejo (1998).

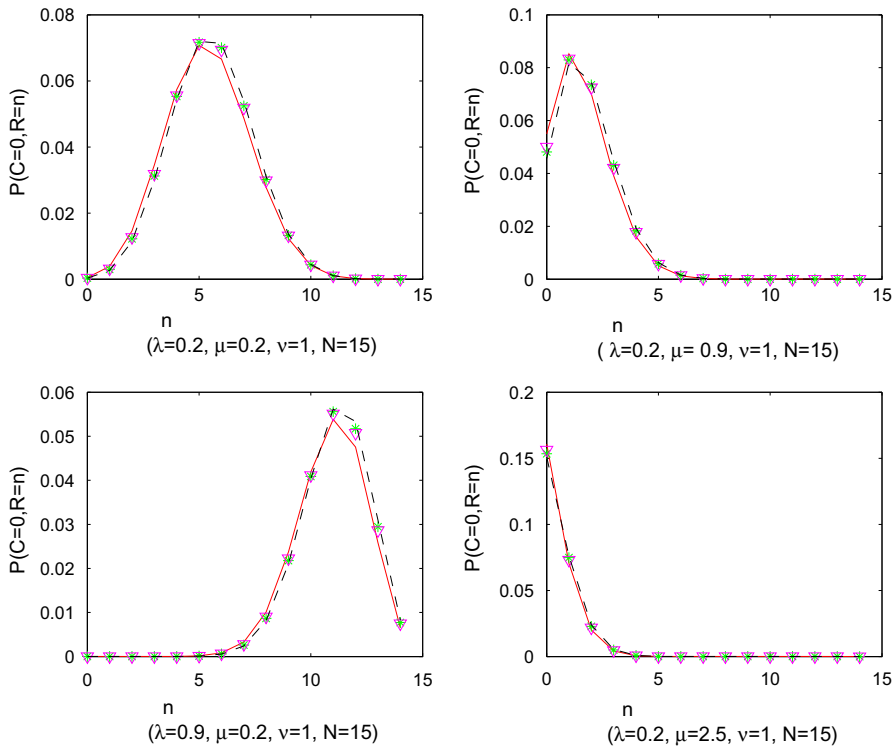


Fig. 1 Distribution of the orbit size when the server is idle

Figures 3 and 4 have the same structure as Fig. 2, but refer to the mean orbit size, $E[R]$ and the mean rate of call generation, $\bar{\lambda}$. Because of the specific behaviour of the blocking probability, P_B , there are two figures, concerning this probability. Figure 5 refers only to the system with inactive orbit, while on Fig. 6 we can see the difference between both models: with inactive orbit, presented with solid lines and with retrials, presented with dotted lines (exponential service).

The results, depicted in Figs. 2, 3, 4, 5, and 6 show that:

- In most of the cases the service distribution type has insignificant influence on the considered characteristics. The only exceptions are the dependencies of the mean rate of call generation and of the blocking probability on the parameter μ (Figs. 4, 5, upper right corners).
- The blocking probability, P_B , shows the same behavior as in the finite queues with retrials (Falin and Artalejo 1998; Wang et al. 2011). It has a point of local maximum as a function of the parameter λ . All other dependencies are monotonically increasing or decreasing, most of them in accordance with our intuition. The presented dependencies on μ confirm the results obtained in Sect. 4 and we can see that the limit values are reached even for small values of this parameter.
- The differences between the model with losses and the corresponding model with retrials are significant. In all presented examples the model with inactive customers shows lower values of the mean orbit size (Fig. 3) and the blocking probability (Fig. 6) than the model with retrials. Conversely, the values of the probability that the server is idle (Fig. 2) and

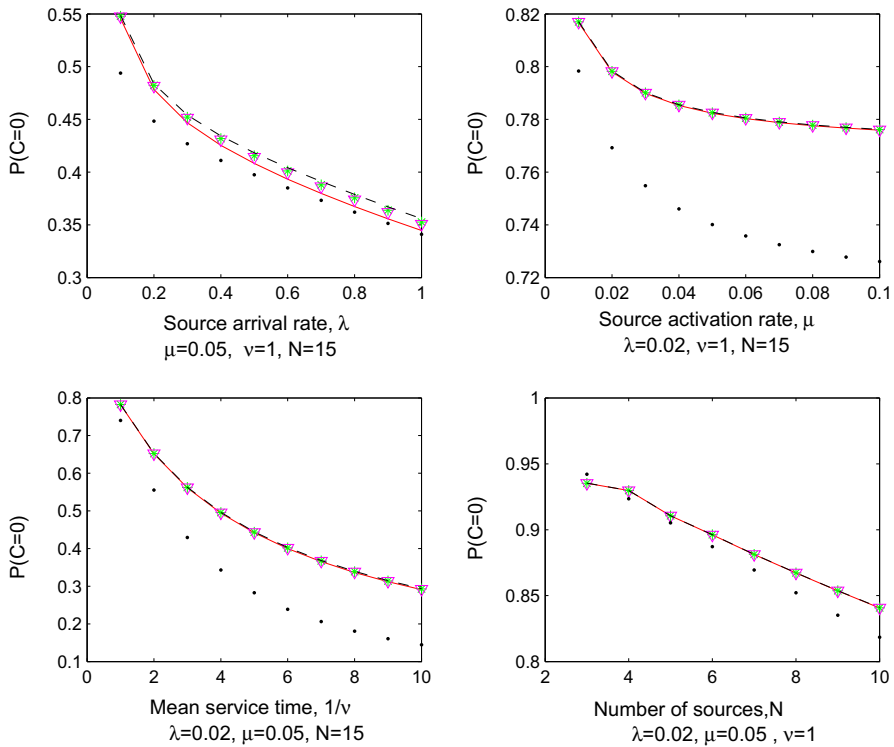


Fig. 2 Probability that the server is idle versus system parameters

the mean rate of call generation (Fig. 4) are higher in the system with losses than in the system with retrials.

A comparison between both models from optimization point of view is considered in the next section.

6 Cost function

In this section we discuss the performance optimization of both models (with losses and with retrials) from a management point of view. To this end we introduce the following cost functions:

$$CF_L = T_S P_0 + T_L E[R],$$

for the system with losses, and

$$CF_R = T_S P_0 + T_R E[R]$$

for the system with retrials. Here T_S, T_R and T_L are positive costs that should be paid for the idle server periods, the delays of the clients and the lost clients, respectively. Our objective is to find optimal values of the system parameters that minimize the cost functions. A theoretical solution of such a problem is almost impossible.

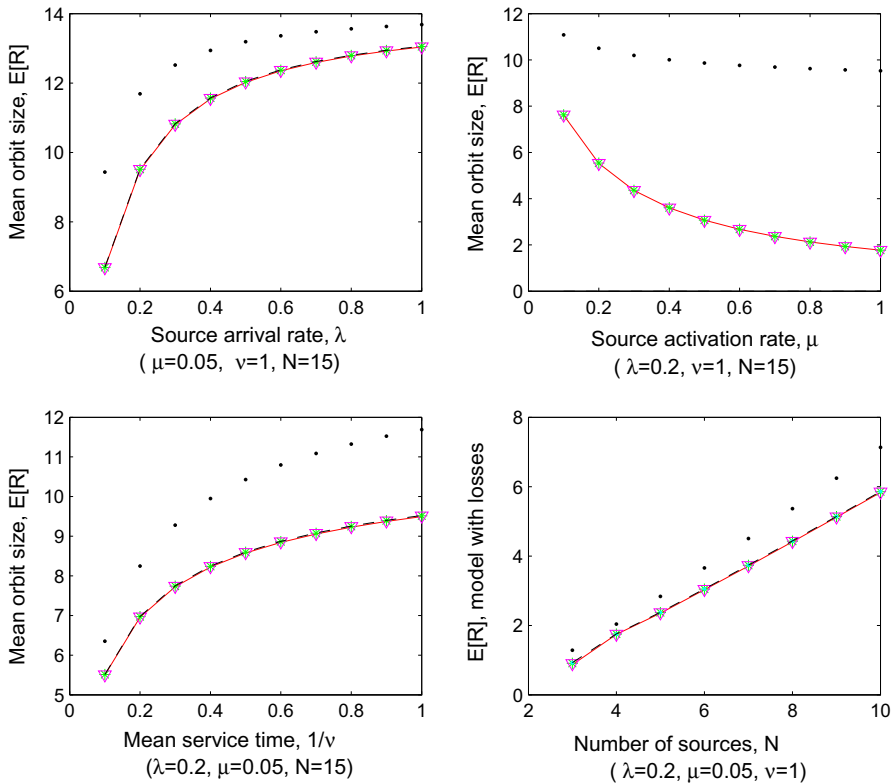


Fig. 3 Mean orbit size versus system parameters

Thus, our aim here is to verify numerically whether the values of CF_R can be minimized by temporarily replacing it with CF_L , i.e. by temporarily blocking all unsuccessful customers for an exponentially distributed time interval. Figures 7 and 8 present the dependence of the function CF_L (the model with losses, solid lines) and the function CF_R (the model with retrials, dotted lines) versus any of the system parameters. We can see that if all costs T_S , T_R and T_L (Fig. 7) are equal to 1, the system with losses shows better results regarding optimization of the system profits. Thus, in this case the system will increase its profits if all unsuccessful clients are lost.

When the cost for a lost client, T_L is greater than the cost for an idle server, T_S and the cost for a delayed client, T_R (Fig. 8), one might intuitively think that the lost clients will decrease system profit. Figure 8 shows that this is not always true, especially for large values of the source arrival rate, λ and small values of the source activation rate, μ .

7 Busy period

Assume that the busy period starts at time $t_0 = 0$ at which all customers are in free state and one of them generates a call. It ends at the first epoch at which the server is free and there are no blocked customers. The length of the busy period is denoted by ζ , its distribution function, $P\{\zeta \leq x\}$ – by $H(x)$ and its Laplace – Stieltjes transform – by $\eta(s)$. For each $t \geq 0$ we consider the following probabilities (densities):

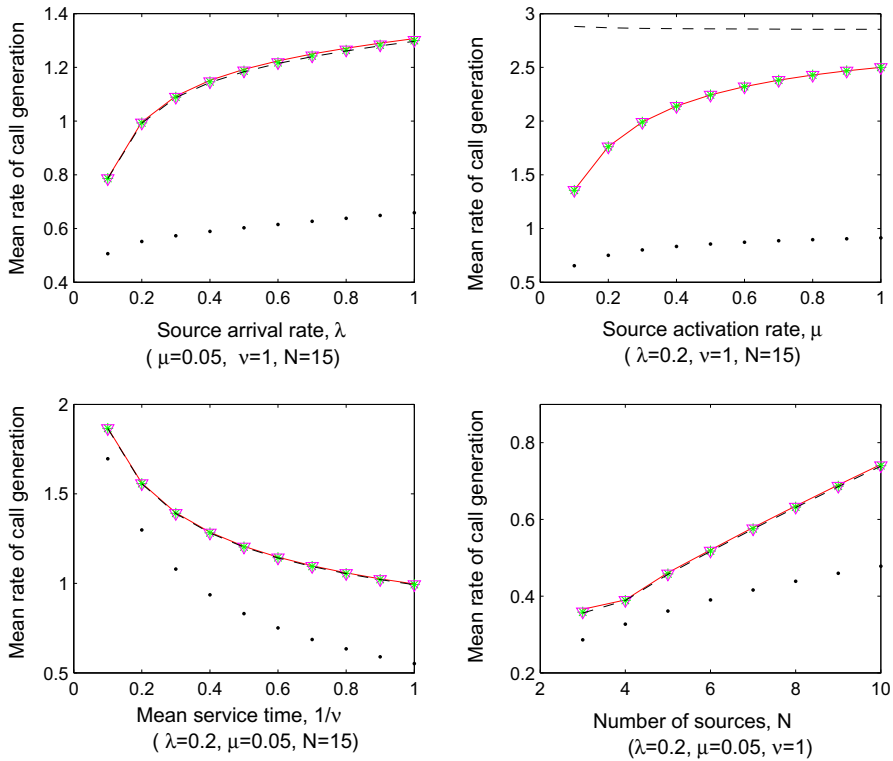


Fig. 4 Mean rate of call generation versus system parameters

$$P_{1n}(t, x)dx = P \{ \zeta > t, C(t) = 1, R(t) = n, x \leq z(t) < x + dx \}, \quad (30)$$

$$0 \leq n \leq N - 1,$$

$$P_{0n}(t) = P \{ \zeta > t, C(t) = 0, R(t) = n \}, 1 \leq n \leq N - 1, \quad (31)$$

with initial conditions

$$P_{0n}(0) = 0, P_{1n}(0, x) = \delta(x)\delta_{0n} \quad (32)$$

and Laplace transforms $\bar{P}_{0n}(s)$ and $\bar{P}_{1n}(s, x)$.

Here $\delta(x)$ is Dirac delta, all other variables are the same as in the previous sections.

The Kolmogorov's equations for these transient probabilities look as follows:

$$\frac{d}{dt} P_{0n}(t) = -[(N - n)\lambda + n\mu]P_{0n}(t) + (n + 1)\mu P_{0,n+1}(t) + \int_0^t P_{1n}(t, x)\gamma(x)dx$$

$$P_{1n}(t, 0) = (N - n)\lambda P_{0n}(t), 1 \leq n \leq N - 1,$$

$$\frac{\partial}{\partial t} P_{1n}(t, x) = - \left[(N - n - 1)\lambda + n\mu + \gamma(x) + \frac{\partial}{\partial x} \right] P_{1n}(t, x)$$

$$+ (n + 1)\mu P_{1,n+1}(t, x) + (N - n)\lambda P_{1,n-1}(t, x), 0 \leq n \leq N - 1,$$

with

$$P_{0N}(t) = P_{1N}(t, x) = P_{1,-1}(t, x) = 0,$$

and initial conditions (32).

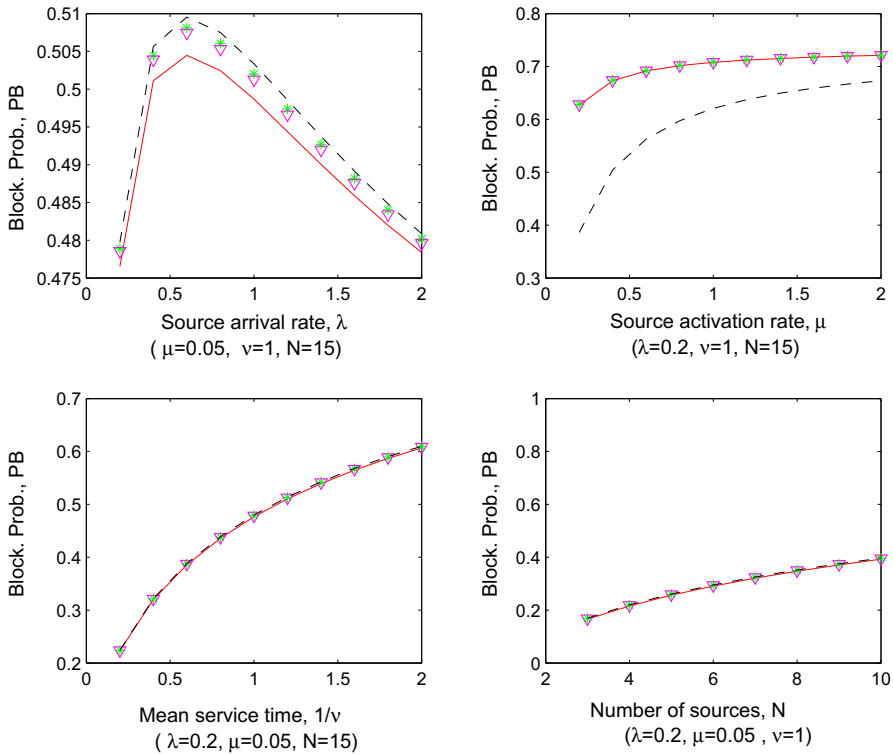


Fig. 5 Blocking probability versus system parameters (in the system with losses)

In addition, the following equalities hold:

$$\frac{d}{dt}H(t) = \int_0^\infty P_{10}(t, x)\gamma(x)dx + \mu P_{01}(t),$$

$$\sum_{n=1}^{N-1} P_{0n}(t) + \sum_{n=0}^{N-1} \int_0^\infty P_{1n}(t, x)dx = 1 - H(t).$$

Applying Laplace transforms in these equations, we get

$$[(N - n)\lambda + n\mu + s]\bar{P}_{0n}(s) = (n + 1)\mu\bar{P}_{0,n+1}(s) + \int_0^\infty \bar{P}_{1n}(s, x)\gamma(x)dx, \quad (33)$$

$$\bar{P}_{1n}(s, 0) = (N - n)\lambda\bar{P}_{0n}(s), \quad 1 \leq n \leq N - 1, \quad (34)$$

$$\left[(N - n - 1)\lambda + n\mu + \gamma(x) + s + \frac{\partial}{\partial x} \right] \bar{P}_{1n}(s, x) = \delta(x)\delta_{n0} + (n + 1)\mu\bar{P}_{1,n+1}(s, x) + (N - n)\lambda\bar{P}_{1,n-1}(s, x), \quad 0 \leq n \leq N - 1, \quad (35)$$

$$\eta(s) = \int_0^\infty \bar{P}_{10}(s, x)\gamma(x)dx + \mu\bar{P}_{01}(s), \quad (36)$$

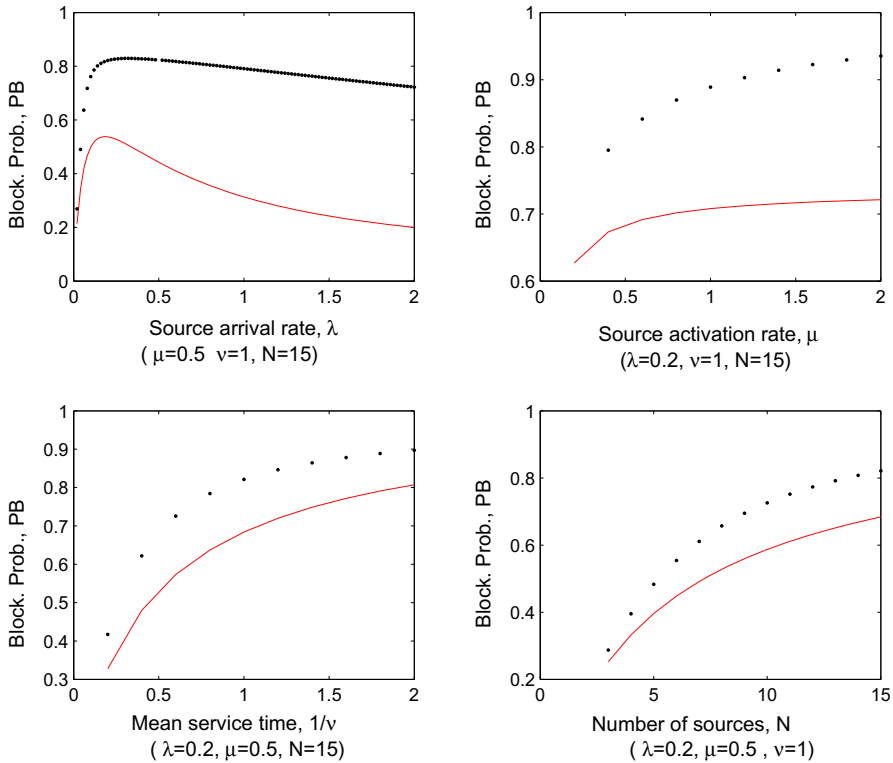


Fig. 6 Blocking probability versus system parameters (in both systems: with losses and with retrials)

$$\sum_{n=1}^{N-1} \bar{P}_{0n}(s) + \sum_{n=0}^{N-1} \int_0^\infty \bar{P}_{1n}(s, x) dx = \frac{1 - \eta(s)}{s}, \quad s \neq 0. \tag{37}$$

According to the discrete transformations method, we write Eq. (35) in a matrix form

$$[\tilde{\theta}I - A] \bar{P}_1(s, x) = D(x),$$

where:

$$\tilde{\theta} = \gamma(x) + s + \frac{\partial}{\partial x},$$

I is the identity matrix of order N ,

$$\begin{aligned} \bar{P}_1(s, x) &= (\bar{P}_{10}(s, x), \dots, \bar{P}_{1,N-1}(s, x))^T, \\ D(x) &= (\delta(x), 0, \dots, 0)^T, \end{aligned}$$

A is the matrix constructed from (35) in the usual way and is the same as in Eq. (7), Sect. 3. Thus, the discrete transformation converting (35) into a simpler form is the same as in Sect. 3. So, the transformation

$$\bar{P}_1(s, x) = Y \bar{Q}_1(s, x) \tag{38}$$

converts (35) into the form

$$[\tilde{\theta}I - A] \bar{Q}_1(s, x) = Y^{-1} D(x). \tag{39}$$

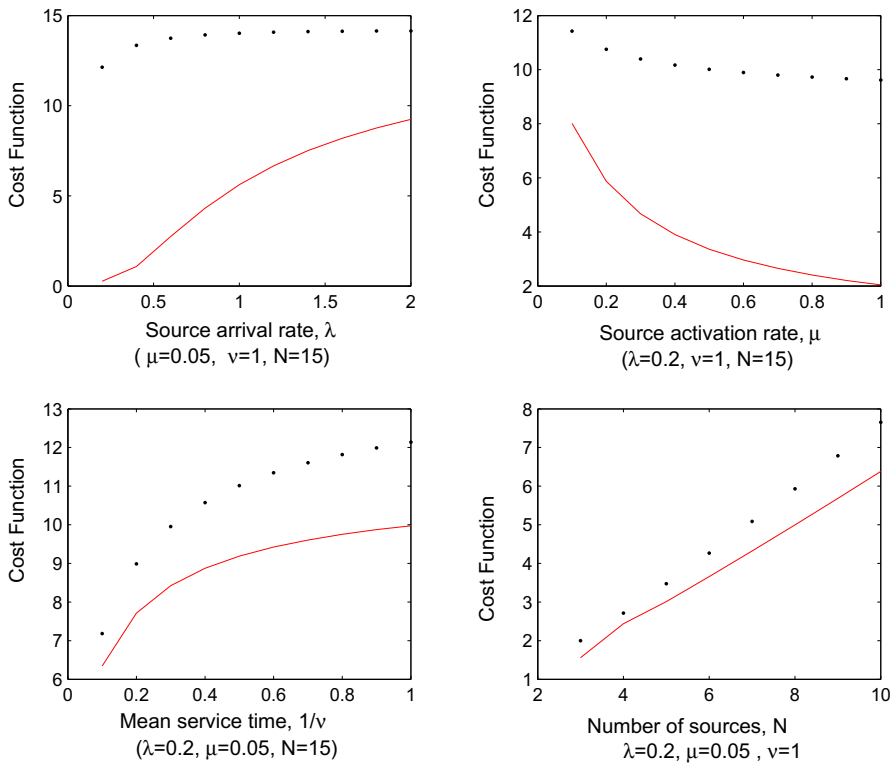


Fig. 7 Cost functions for $T_S = T_R = T_L = 1$

Because only the first coordinate of the vector $D(x)$ is nonzero, it is sufficient to find only the first column of the matrix Y^{-1} , $(\bar{y}_0^{(0)}, \dots, \bar{y}_{N-1}^{(0)})^T$. For its coordinates we found:

$$\bar{y}_k^{(0)} = \binom{N-1}{k} \left(\frac{\lambda}{\mu}\right)^k \left(\frac{\mu}{\lambda+\mu}\right)^{N-1-k}. \tag{40}$$

Equation (39) and relation (38) allow to express the functions $\bar{P}_{1n}(s, x)$ in terms of N unknown quantities, the initial values $\bar{Q}_{1n}(s, 0)$. Then, from (34) and (36) we express $\bar{P}_{0n}(s)$ and $\eta(s)$ in terms of the same unknowns, $\bar{Q}_{1n}(s, 0)$. Finally, substituting in (33) and (37) we derive a system of linear equations for $\bar{Q}_{1n}(s, 0)$. Thus, we can prove the following theorem.

Theorem 2 *The Laplace – Stieltjes transform, $\eta(s)$, of the distribution function of the busy period length can be calculated by the formula*

$$\eta(s) = \sum_{k=0}^{N-1} \bar{Q}_{1k}(s, 0) \left[1 + g_k(s) - \frac{k(\lambda + \mu)}{(N-1)\lambda} \right] + \sum_{k=0}^{N-1} g_k(s) \bar{y}_k^{(0)}, \tag{41}$$

where $\bar{Q}_{1k}(s, 0)$ satisfy the following system of linear equations:

$$\sum_{k=0}^{N-1} \bar{Q}_{1k}(s, 0) \left[y_n^{(k)} \left(\frac{s + n\mu}{(N-n)\lambda} - g_k(s) \right) - \frac{k(\lambda + \mu)}{(N-n-1)\lambda} \sum_{i=0}^n y_i^{(k)} \right]$$

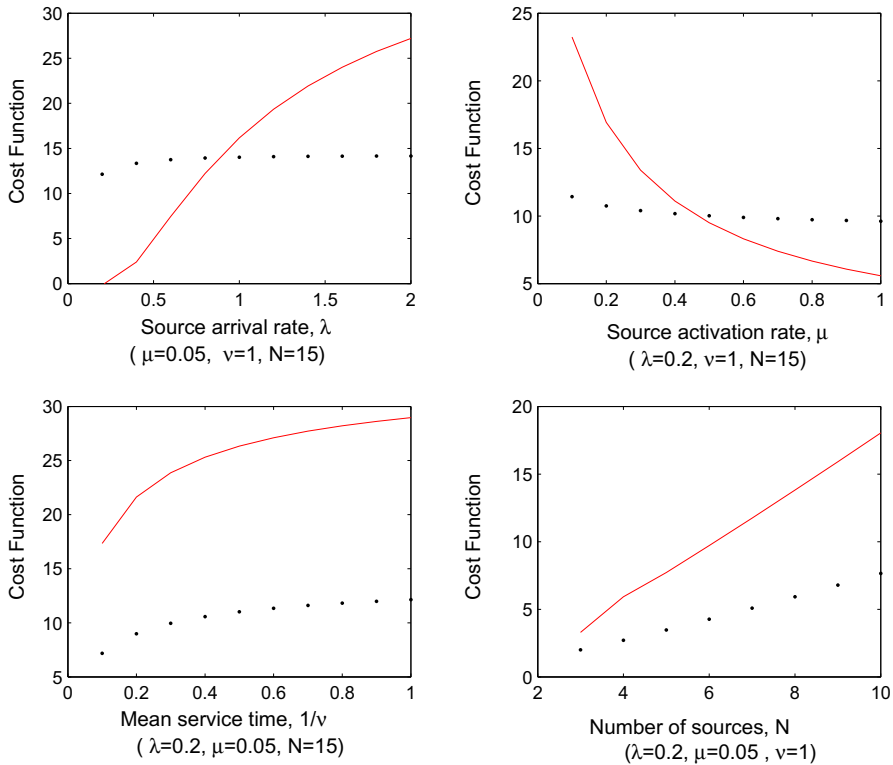


Fig. 8 Cost functions for $T_S = T_R = 1, T_L = 3$

$$= \sum_{k=0}^{N-1} g_k(s) \bar{y}_k^{(0)} y_n^{(k)}, \quad 1 \leq n \leq N - 1, \tag{42}$$

$$\begin{aligned} & \bar{Q}_{10}(s, 0) \left[(1 - g_0(s)) \left(\frac{\lambda + \mu}{\mu} \right)^{N-1} + \sum_{n=1}^{N-1} \frac{s}{(N - n)\lambda} + 1 + g_0(s) \right] \\ & + \sum_{k=1}^{N-1} \bar{Q}_{1k}(s, 0) \left[s \sum_{n=1}^{N-1} \frac{y_n^{(k)}}{(N - n)\lambda} + 1 + g_k(s) - \frac{k(\lambda + \mu)}{(N - 1)\lambda} \right] \\ & = g_0(s) - \sum_{k=0}^{N-1} g_k(s) \bar{y}_k^{(0)}. \end{aligned} \tag{43}$$

Here $y_n^{(k)}$ are given by the equalities (9)–(12) (Proposition 1), $\bar{y}_n^{(0)}$ —by (40), $g_k(s) = g(k(\lambda + \mu) + s)$.

Thus, to calculate $\eta(s)$ we first use (42) and express all quantities $\bar{Q}_{1k}(s, 0), k = 0, 1, \dots, N - 2$ in terms of the first one— $\bar{Q}_{10}(s, 0)$ (for example, by using Gauss algorithm). Then, substituting consecutively in (43) and (41) we can find $\bar{Q}_{10}(s, 0)$ and $\eta(s)$.

Further, upon suitable differentiations in (41)–(43) we can compute the first moments of the busy period length.

8 Conclusion

This article proposes formulas for computing the joint and marginal steady state distributions of the server state and the orbit size for a single server queueing system, where a customer arriving at a time moment of a busy server is not allowed either to queue or to try again for an exponentially distributed time interval. Such a customer is sent to the orbit of inactive customers. Although the obtained formulas are not as simple as in the corresponding model with retrials, they allow to calculate these distributions and macro characteristics of the system performance. On the basis of numerical results it is shown that, for some values of the system parameters, the performance of the corresponding finite retrial queue can be improved by temporarily forbidding any attempts, if the first one failed. Formulas for computing the Laplace–Stieltjes transform of the distribution function of the busy period length are derived. They are a basis for future investigation of the system in non-stationary regime.

Acknowledgments The author is grateful to the anonymous referees whose precise revisions and useful comments helped much to improve this article.

References

- Amador, J. (2010). On the distribution of the successful and blocked events in retrial queues with finite number of sources. In: *Proceedings of the 5th international conference on Queueing theory and network applications* (pp. 15–22).
- Artalejo, J., & Gómez-Corral, A. (2007). Modeling communication systems with phase type service and retrial times. *IEEE Communications Letters*, *11*, 955–957.
- Artalejo, J., & Gómez-Corral, A. (2008). *Retrial queueing systems: A computational approach*. Berlin: Springer.
- Artalejo, J., & Phung-Duc, T. (2012). Markovian retrial queues with two way communications. *Journal of Industrial and Management Optimization*, *8*(4), 781–806.
- Dahmouni, H., Morin, B., & Vaton, S. (2005). Performance modeling for GSM/GPRS cells with different radio resource allocation strategies. *IEEE Proceedings of Wireless Communications and Networking Conference*, *3*, 1317–1322.
- De Kok, A. (1984). Algorithmic methods for single server system with repeated attempts. *Statistica Neerlandica*, *38*, 23–32.
- Dragieva, V. (2013). A finite source retrial queue: Number of retrials. *Communications in Statistics - Theory and Methods*, *42*(5), 812–829.
- Drekic, S., & Grassmann, W. (2002). An eigenvalue approach to analyzing a finite source priority queueing model. *Annals of Operations Research*, *112*, 139–152.
- Efrosinin, D., & Breuer, L. (2006). Threshold policies for controlled retrial queues with heterogeneous servers. *Annals of Operations Research*, *141*, 139–162.
- Falin, G., & Artalejo, J. (1998). A finite source retrial queue. *European Journal of Operational Research*, *108*, 409–424.
- Falin, G., & Templeton, J. (1997). *Retrial Queues*. London: Chapman and Hall.
- Jaiswal, N. (1969). *Priority queues*. New York: Academic Press.
- Janssens, G. (1997). The quasi-random input queueing system with repeated attempts as a model for collision-avoidance star local area network. *IEEE Transactions on Communications*, *45*, 360–364.
- Kim, Ch., Klimenok, V., Birukov, A., & Dudin, A. (2006). Optimal multi-threshold control by the BMAP/SM/1 retrial system. *Annals of Operations Research*, *141*, 193–210.
- Lee, H., Moon, J., Kim, B., Park, J., & Lee, S. (2005). A simple eigenvalue method for low-order D-BMAP/G/1 queues. *Applied Mathematical Modelling*, *29*, 277–288.
- Li, H., & Yang, T. (1995). A single server retrial queue with server vacations and a finite number of input sources. *European Operational Research*, *85*, 149–160.
- Mehmet-Ali, M., Elhakeem, A., & Hayes, J. (1988). Traffic analysis of a local area network with star topology. *IEEE Transactions on Communications*, *36*, 703–712.
- Moscholios, L., & Logothetis, M. (2006). Engset multi-rate state-dependent loss models with QoS guarantee. *International Journal of Communications Systems*, *19*(1), 67–93.

- Ohmura, H., & Takahashi, Y. (1985). An analysis of repeated call model with a finite number of sources. *Electronics and Communications in Japan*, 68, 112–121.
- Overby, H. (2005). Performance modelling of optical packet switched networks with the Engset traffic model. *Optics Express*, 13, 1685–1695.
- Tran-Gia, P., & Mandjes, M. (1997). Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE Journal on Selected Areas in Communications*, 15, 1406–1414.
- Van Do, T., Wochner, P., Berches, T., & Sztrik, J. (2014). A new finite-source queueing model for mobile cellular networks applying spectrum renting. *Asia-Pacific Journal of Operational Research*, 31, 14400004_19.
- Wang, J., Zhao, L., & Zhang, F. (2011). Analysis of the finite source retrial queues with server breakdowns and repairs. *Journal of Industrial and Management Optimization*, 7, 655–676.
- Zhang, F., & Wang, J. (2013). Performance analysis of the retrial queues with finite number of sources and service interruptions. *Journal of the Korean Statistical Society*, 42(1), 117–132.
- Zukerman, M., Wong, E., Rosberg, Z., Myoung, G., & Le Vu, H. (2004). On teletraffic applications to OBS. *IEEE Communications Letters*, 8, 116–118.