Annals of Operations Research manuscript No.
(will be inserted by the editor)

A matrix continued fraction approach to multiserver retrial queues

Tuan Phung-Duc · Hiroyuki Masuyama ·
Shoji Kasahara · Yutaka Takahashi

Received: date / Accepted: date

Abstract We consider basic M/M/c/c ($c \geq 1$) retrial queues where the number of busy servers and that of customers in the orbit form a level-dependent quasi-birth-and-death (QBD) process with a special structure. Based on this structure and a matrix continued fraction approach, we develop an efficient algorithm to compute the joint stationary distribution of the numbers of busy servers and retrial customers. Through numerical experiments, we demonstrate that our algorithm works well even for M/M/c/c retrial queues with large value of c.

Keywords Multiserver retrial queue · Level-dependent QBD · Matrix continued fraction · Call center · Resource dimensioning · System planning

Mathematics Subject Classification (2000) 68M20 · 90B22 · 60K25

1 Introduction

This paper considers M/M/c/c retrial queues, in which if an arriving customer finds an idle server, he starts to be served, otherwise he moves to a virtual orbit, stays there for an exponentially distributed time and retries to get service. Retrial queues arise in various systems such as telecommunications, computer networks and call centers [1,6,7, 21,22,27,31]. Aguir et al. [1] investigate the impact of retrials on the performance of call centers, using a fluid approximation. Artalejo and Pla [6] further evaluate the effect of customer retrials on operations of telecommunication systems, by a retrial queue with

Tuan Phung-Duc · Hiroyuki Masuyama · Shoji Kasahara · Yutaka Takahashi
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 JAPAN
Tel.: +81-75-753-5879 Fax: +81-75-753-3358
E-mail: tuan@sys.i.kyoto-u.ac.jp

masuyama@sys.i.kyoto-u.ac.jp · shoji@i.kyoto-u.ac.jp · takahashi@i.kyoto-u.ac.jp

infinite waiting room and orbit. The authors in [6] propose two truncation methods to analyze the underlying Markov chain of the retrial queue. Koole and Mandelbaum [21], and Gans et al. [22] present extensive surveys on queueing models of call centers in which the retrial phenomenon is taken into account.

Considerable attention has been paid to applications of retrial queueing models for the performance evaluation of cellular mobile and computer networks [2,7,12,24, 27,31]. Tran-Gia and Mandjes [31] use some retrial queueing models to analyze the performance of cellular mobile networks and claim that retrial phenomenon should be taken into account in a careful design of these systems. Marsan et al. [24] propose an approximation for the blocking probability of a retrial queue presented in [31]. Alfa and Li [2] and Choi et al. [12] consider the design and analysis of cellular mobile networks with correlated arrival processes. Artalejo and Lopez-Herrero [7] evaluate the performance of cellular mobile networks operating under a random environment. Phung-Duc et al. [27] develop a multiserver retrial queueing model with random server selection to study the influence of retransmissions due to the contention of bursts on Optical Burst Switched Networks (OBS) with wavelength conversion.

Analytical solutions for multiserver retrial queues have been obtained for a few special cases. An explicit solution for the joint stationary distribution of the numbers of busy servers and customers in the orbit is obtained only for the M/M/1/1 retrial queue. For the case of $c = 2$, the joint stationary distribution is expressed in terms of hypergeometric functions [14,18,29]. As for the cases of $c = 3$ and 4, Phung-Duc et al. [28] show that the joint stationary distribution is expressed in terms of continued fractions. The same authors in [29] further derive analytical solutions for the joint stationary distribution of state-dependent M/M/c/$c + r$ retrial queues with Bernoulli abandonment, where $c + r \leq 4$.

For general M/M/c/c retrial queues, many approximation methods have been developped so far. The basic idea of these methods is that the original M/M/c/c retrial queue is approximated by some other analytically tractable models. A direct truncation method for the M/M/c/c retrial queue, assumes that the number of customers in the orbit does not exceed some truncation point, under which a blocked customer that sees the orbit full is lost [14]. Stepanov [32] considers some truncation methods which disregard the states whose stationary probabilities are considered to be small. Falin and Templeton [14] introduce a generalized truncation method which assumes that all the servers are always busy due to retrials from the orbit when the number of customers in the orbit exceeds some sufficiently large level.

Artalejo and Pozo [4] propose an extension of [14] assuming that there is at most one idle server when the number of customers in the orbit exceeds some level. The authors claim that it is difficult to make a further extension due to the same difficulty as in the derivation of an analytical solution for the M/M/c/c retrial queue with $c > 2$. It should be noted that in these truncation methods [4,14], the number of retrial customers is not necessarily assumed to be finite. Anisimov and Artalejo [3] present a unified approach to prove that the stationary distributions of these generalized truncation models converge to those of the original models.

Matrix analytic approaches to mutiserver retrial queues are extensively studied by many researchers. For a detailed list of papers on this research direction, the readers are referred to a survey paper by Gomez-Corral [17] and a recently published book by Altalejo and Gomez-Corral [5]. Breuer et al. [9] and Klimenok and Dudin [20] consider a BMAP/PH/N retrial queue, which is more general than the model of this paper. The authors formulate the dynamics of the queue as a level-dependent M/G/1 type Markov process with the so-called quasi-Toeplitz structure. The computational algorithm in [9, 20] is based on G-matrices, which do not have a sparse structure.

Neuts and Rao [25] propose an approximation method which assumes that the retrial rate is constant when the number of customers in the orbit exceeds a *truncation point*. Under this assumption, the level-dependent QBD process becomes a level-independent QBD process from the truncation point. As a result, the authors obtain a level-independent QBD process with multiple boundary conditions for which some efficient algorithms are available. The authors show that their approximation outperforms the direct truncation method in [14] when the traffic intensity is low. However, under a high traffic intensity the approximation by Neuts and Rao [25] has a large error because the dynamics of the approximation model is changed. Domenech-Benlloch et al. [13] improve the method of Neuts and Rao by adjusting the retrial rate.

Hanschke [19] analyzes a mutiserver retrial queue, in which arriving customers are blocked with some positive probabilities depending on the number of busy servers. In this retrial queue, the number of busy servers and that of customers in the orbit form a level-dependent QBD process with some special structure, which enables us to calculate the rate matrices by a forward-type algorithm. However, it is reported by Baumann and Sandmann [8] that the forward type algorithm in [19] is numerically unstable due to the mix of positive and negative terms in calculation. Furthermore, the approach in [19] cannot be applied to the M/M/c/c retrial queue of this paper. It should be noted that the approach by Hanschke [19] is different from those presented in [4,14,25,32] in the sense that the author directly analyzes the original retrial queue.

Liu and Zhao [23] use a censoring technique and a level-dependent QBD approach to derive analytical solutions for the cases of $c = 1$ and 2 and show the asymptotic behavior for the stationary distribution of the general case. Bright and Taylor [10] develop a computational algorithm for the rate matrices and the stationary distributions of level-dependent QBD Markov processes, which can be used to analyze the M/M/c/c retrial queue. The authors also propose a method to determine the truncation level, however, the method unfortunately cannot be applied for the level-dependent QBD process arising from the M/M/c/c retrial queue. It should be noted that an efficient estimation of the truncation point for level-dependent QBD processes plays a crucial role in the computation of their stationary distributions.

Recently, some progress has been made in the development of computational algorithms for level-dependent QBD processes. Baumann and Sandmann [8] propose a backward algorithm for level-dependent QBD processes. Through numerical experiments, the authors show that their algorithm outperforms several conventional numerical methods such as Gaussian elimination method, Jacobi iteration method, Gauss-Seidel iteration method, and the power method on uniformized discrete-time level-dependent QBDs. However, in [8] a comparison with the most competitive algorithm by Bright and Taylor [10] has not been carried out yet. Phung-Duc et al. [30] independently develop a similar algorithm to that of Baumann and Sandmann [8] for level-dependent QBD processes. The authors in [30] theoretically show that their algorithm outperforms the algorithm by Bright and Taylor [10] in memory usage, while the computational complexities of both algorithms are the same.

In this paper, we propose an efficient algorithm to compute the joint stationary distribution of the M/M/c/c retrial queue, based on the backward algorithms presented in [8,30]. First, using a special structure of the QBD process, we show that only the last row vector of the rate matrices is nonzero and thus the computation of the rate matrices is reduced to that of their last row vectors. Second, we propose an algorithm to compute the last row vectors efficiently in both memory usage and computational complexity. A remarkable feature of the proposed algorithm is that it does not require the computation of any inverse matrix.

Furthermore, we use the analytical result of an M/M/1/1 retrial queue to determine the truncation point for the level-dependent QBD process of the M/M/c/c retrial queue. Using this truncation point, we compute a numerical solution for the joint stationary distribution of the M/M/c/c retrial queue. It should be noted that the truncation methods in [4,14,23,25] aim at minimizing the truncation point. In contrast, our truncation method aims at finding a sufficiently large truncation point. In addition, the same as Hanschke [19], we also directly analyze the original M/M/c/c retrial queues. An important remark is that the computational complexity has not been shown in the literature [13,19,25] on matrix analytic approaches to multiserver retrial queues. In this paper, we show that the computational complexity of the algorithm is linear with respect to the number of servers. We further show that our algorithm is numerically stable because it manipulates positive numbers.

The rest of the paper is organized as follows. Section 2 presents the M/M/c/c retrial queueing model and some preliminary results. The main contribution of this paper is presented in Section 3. Section 4 is devoted to an extensive presentation of numerical examples for various scenarios. Finally, Section 5 concludes the paper.

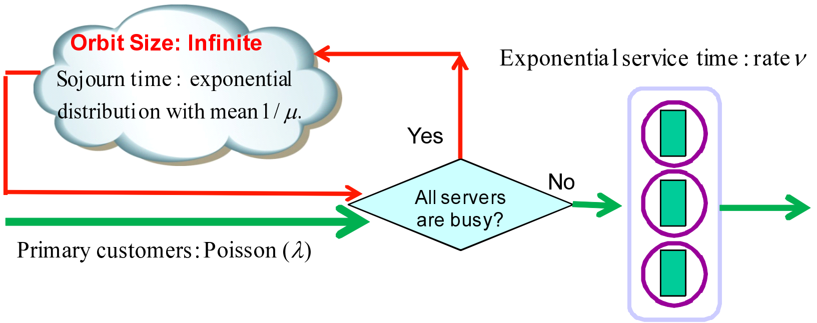
Orbit Size: Infinite
Sojourn time : exponential distribution with mean 1 / μ.
Exponential service time : rate ν
Yes
No
All servers are busy?
Primary customers : Poisson (λ)

Fig. 1 M/M/c/c retrial queue.

2 Model Description and Preliminary Results

We first describe the M/M/c/c retrial queue. The primary customers arrive at the servers according to a Poisson process with rate $\lambda > 0$ and the service time of each server follows an exponential distribution with mean $1/\nu$. An arriving primary customer either occupies one of idle servers if any or moves to the orbit if all the servers are busy. A customer in the orbit is called a *retrial customer* hereafter. Each retrial customer stays in the orbit for an exponentially distributed time with finite positive mean $1/\mu$ independently of other customers. After the sojourn time in the orbit, a retrial customer retries to get service. The retrial customer is served immediately if there is an idle server upon arrival, otherwise it joins the orbit again. See Fig. 1 for details.

Let $X(t) = (C(t), N(t))$ $(t \geq 0)$, where $C(t)$ and $N(t)$ denote the numbers of busy servers and customers in the orbit, at time t, respectively. It is easy to see that the bivariate process $\{X(t); t \geq 0\}$ is a Markov chain with the state space $\{0, 1, \dots, c\} \times \mathbb{Z}_+$, where $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$. Throughout the paper, we assume that $\{X(t)\}$ is ergodic. It is shown in the book by Falin and Templeton [14] that the necessary and sufficient condition for the ergodicity of $\{X(t)\}$ is $\rho = \lambda/(c\nu) < 1$.

It is easy to confirm that $\{X(t)\}$ is a level-dependent QBD process whose infinitesimal generator is given by

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_1^{(0)} & \boldsymbol{Q}_0^{(0)} & \boldsymbol{O} & \boldsymbol{O} & \cdots \\ \boldsymbol{Q}_2^{(1)} & \boldsymbol{Q}_1^{(1)} & \boldsymbol{Q}_0^{(1)} & \boldsymbol{O} & \cdots \\ \boldsymbol{O} & \boldsymbol{Q}_2^{(2)} & \boldsymbol{Q}_1^{(2)} & \boldsymbol{Q}_0^{(2)} & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{Q}_2^{(3)} & \boldsymbol{Q}_1^{(3)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $\boldsymbol{O}$ denotes a matrix of an appropriate dimension with entries being zeros and $\boldsymbol{Q}_0^{(n)}$, $\boldsymbol{Q}_1^{(n)}$ and $\boldsymbol{Q}_2^{(n)}$ $(n \in \mathbb{Z}_+)$ are given by

$$\boldsymbol{Q}_0^{(n)} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}, \qquad \boldsymbol{Q}_2^{(n)} = \begin{pmatrix} 0 & n\mu & 0 & \cdots & 0 \\ 0 & 0 & n\mu & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & 0 & n\mu \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix},$$

$$\boldsymbol{Q}_1^{(n)} = \begin{pmatrix} b_0^{(n)} & \lambda & 0 & \cdots & \cdots & 0 \\ \nu & b_1^{(n)} & \lambda & \ddots & & \vdots \\ 0 & 2\nu & b_2^{(n)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & b_{c-1}^{(n)} & \lambda \\ 0 & \cdots & \cdots & 0 & c\nu & b_c^{(n)} \end{pmatrix}.$$

The diagonal components of $\boldsymbol{Q}_1^{(n)}$ are given by $b_i^{(n)} = -(\lambda + i\nu + n\mu(1 - \delta_{i,c}))$ for $i = 0, 1, \ldots, c$, where $\delta_{i,c}$ denotes the Kronecker delta. Let $\pi_{i,n} = \lim_{t\to\infty} \Pr\{C(t) = i, N(t) = n\}$ denote the joint stationary probability of the numbers of busy servers and customers in the orbit. Let $\boldsymbol{\pi}_n = (\pi_{0,n}, \pi_{1,n}, \ldots, \pi_{c,n})$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$. The stationary distribution $\boldsymbol{\pi}$ is the solution of the following system of equations.

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

where vectors $\boldsymbol{e}$ and $\boldsymbol{0}$ denote a column vector and a row vector with an appropriate dimension whose entries are ones and zeros, respectively. Equation (1) is rewritten in a vector form as follows.

$$\boldsymbol{\pi}_{n-1}\boldsymbol{Q}_0^{(n-1)} + \boldsymbol{\pi}_n\boldsymbol{Q}_1^{(n)} + \boldsymbol{\pi}_{n+1}\boldsymbol{Q}_2^{(n+1)} = \boldsymbol{0}, \qquad n \in \mathbb{N}, \tag{2}$$

$$\boldsymbol{\pi}\boldsymbol{e} = 1, \tag{3}$$

where $\mathbb{N} = \{1, 2, \ldots\}$. The solution of (2) and (3) is given by

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1}\boldsymbol{R}^{(n)}, \qquad n \in \mathbb{N},$$

where $\{\boldsymbol{R}^{(n)}; n \in \mathbb{N}\}$ is the minimal nonnegative solution of

$$\boldsymbol{Q}_0^{(n-1)} + \boldsymbol{R}^{(n)}\boldsymbol{Q}_1^{(n)} + \boldsymbol{R}^{(n)}\boldsymbol{R}^{(n+1)}\boldsymbol{Q}_2^{(n+1)} = \boldsymbol{O}, \qquad n \in \mathbb{N},$$

and the boundary vector $\boldsymbol{\pi}_0$ is the solution of

$$\boldsymbol{\pi}_0(\boldsymbol{Q}_1^{(0)} + \boldsymbol{R}^{(1)}\boldsymbol{Q}_2^{(1)}) = \boldsymbol{0},$$
$$\boldsymbol{\pi}_0(\boldsymbol{I} + \boldsymbol{R}^{(1)} + \boldsymbol{R}^{(1)}\boldsymbol{R}^{(2)} + \ldots)\boldsymbol{e} = 1.$$

Matrix $\boldsymbol{I}$ denotes an identity matrix with an appropriate dimension.

Proposition 1 *We have the following backward recursive equation.*

$$\boldsymbol{R}^{(n)} = R_n(\boldsymbol{R}^{(n+1)}), \qquad n \in \mathbb{N},$$

where $R_n(\boldsymbol{X})$ *is defined as*

$$\begin{aligned} &R_n : M \longrightarrow M, \\ &R_n(\boldsymbol{X}) \;=\; \boldsymbol{Q}_0^{(n-1)}\left(-\boldsymbol{Q}_1^{(n)} - \boldsymbol{X}\boldsymbol{Q}_2^{(n+1)}\right)^{-1}, \qquad n \in \mathbb{N}. \end{aligned}$$

Here, M denotes a set of $(c+1)\times(c+1)$ matrices in which $R_n(\cdot)$ is well defined.

Definition 1 For $k, n \in \mathbb{N}$, we define the sequence $\boldsymbol{R}_k^{(n)}$ as follows.

$$\boldsymbol{R}_0^{(n)} = \boldsymbol{O}, \qquad n \in \mathbb{N},$$

and

$$\boldsymbol{R}_k^{(n)} = R_n(\boldsymbol{R}_{k-1}^{(n+1)}) = \cdots = R_n \circ R_{n+1} \circ \circ \circ R_{n+k-1}(\boldsymbol{O}), \qquad n,\ k \in \mathbb{N}, \tag{4}$$

where $f \circ g(\cdot) = f(g(\cdot))$.

Proposition 2 (Proposition 2.4 in [30]) *For $k, n \in \mathbb{N}$, we have*

$$\lim_{k\to\infty} \boldsymbol{R}_k^{(n)} = \boldsymbol{R}^{(n)}.$$

Remark 1 According to Proposition 1 and (4), $\boldsymbol{R}^{(n)}$ can be regarded as an infinite matrix continued fractions, and $\boldsymbol{R}_k^{(n)}$ is the k-th approximation of $\boldsymbol{R}^{(n)}$. For the general notion of continued fractions, the readers are referred to [15,16,26].

3 Computation of the Stationary Distribution

In this section, we develop an algorithm to compute an approximation to the stationary distribution of the level-dependent QBD process arising from the M/M/c/c retrial queue. The algorithm is divided in three stages. The first stage is devoted to the computation of a fundamental step of the backward formula (4). The second stage is concerned with the computation of an approximation to the stationary distribution of the level-dependent QBD process, provided that the truncation point is given in advance. The final stage presents a method to determine the truncation point, based on the explicit solution of an M/M/1/1 retrial queue.

3.1 Fundamental step

According to Definition 1, we have to compute k inverse matrices in order to obtain $\boldsymbol{R}_k^{(n)}$. The computational cost for each inverse matrix is equal to $O(c^3)$, where $f(x) = O(x^n)$ $(n \in \mathbb{N})$ implies that there exists some finite $\kappa > 0$ such that $\lim_{x\to\infty} |f(x)|/x^n = \kappa$. Instead of directly computing the inverse matrix, we propose an efficient method to obtain $\boldsymbol{R}_k^{(n)}$ and an approximation to $\boldsymbol{R}^{(n)}$, based on a sparse structure of these rate matrices. The computational complexity of the proposed method is only $O(c)$. Indeed, the first c rows of $\boldsymbol{R}_k^{(n)}$ and $\boldsymbol{R}^{(n)}$ are all zeros, due to the special structure of $\boldsymbol{Q}_0^{(n-1)}$. Therefore, the computation of $\boldsymbol{R}_k^{(n)}$ and $\boldsymbol{R}^{(n)}$ is reduced to that of $\boldsymbol{r}_k^{(n)}$ and $\boldsymbol{r}^{(n)}$, where $\boldsymbol{r}_k^{(n)}(k \in \mathbb{Z}_+)$ and $\boldsymbol{r}^{(n)}$ denote the last rows of $\boldsymbol{R}_k^{(n)}$ and $\boldsymbol{R}^{(n)}$, respectively. Let

$$\boldsymbol{r}_k^{(n)} = \left(r_{k,0}^{(n)}, r_{k,1}^{(n)}, \dots, r_{k,c}^{(n)}\right), \qquad \boldsymbol{r}^{(n)} = \left(r_0^{(n)}, r_1^{(n)}, \dots, r_c^{(n)}\right), \qquad n,\ k \in \mathbb{N}.$$

Remark 2 The sparse structure of the rate matrices is also used by Liu and Zhao [23], who exploit the special structure of $\boldsymbol{R}^{(n)}$ to derive explicit solutions for the M/M/c/c retrial queues with $c = 1, 2$, and some asymptotic results for the general case. In this paper the sparse structure of the rate matrices is used in order to reduce the computational complexity of a numerical algorithm.

Theorem 1 *For $n,\ k \in \mathbb{N}$, $\boldsymbol{r}_k^{(n)}$ is expressed in terms of $\boldsymbol{r}_{k-1}^{(n+1)}$ by*

$$r_{k,i}^{(n)} = \alpha_i + \beta_i r_{k,c}^{(n)}, \qquad i = 0, 1, \dots, c-1,$$

where $\{\alpha_i, \beta_i; i = c-1, c-2, \dots, 1\}$ and $r_{k,c}^{(n)}$ are determined as follows.

$$\alpha_c = 0, \quad \beta_c = 1, \quad \alpha_{c-1} = -1, \quad \beta_{c-1} = -\frac{b_c^{(n)} + (n+1)\mu r_{k-1,c-1}^{(n+1)}}{\lambda},$$

$$\alpha_{i-1} = -\frac{b_i^{(n)}\alpha_i + (i+1)\nu\alpha_{i+1}}{\lambda}, \quad i = c-1, c-2, \dots, 1$$

$$\beta_{i-1} = -\frac{b_i^{(n)}\beta_i + (i+1)\nu\beta_{i+1} + (n+1)\mu r_{k-1,i-1}^{(n+1)}}{\lambda}, \quad i = c-1, c-2, \dots, 1$$

and

$$r_{k,c}^{(n)} = -\frac{b_0^{(n)}\alpha_0 + \nu\alpha_1}{b_0^{(n)}\beta_0 + \nu\beta_1}.$$

Proof Let $\boldsymbol{U}_k^{(n)}$ denote

$$\boldsymbol{U}_k^{(n)} = \boldsymbol{Q}_1^{(n)} + \boldsymbol{R}_{k-1}^{(n+1)}\boldsymbol{Q}_2^{(n+1)}, \qquad n,\ k \in \mathbb{N}.$$

Matrix $\boldsymbol{U}_k^{(n)}$ is the defective infinitesimal generator of the restricted process of $\{X(t)\}$ on level n, under the taboo of levels $n-1$ and $n+k$. Due to the special structure of $\boldsymbol{R}_{k-1}^{(n+1)}$ and of $\boldsymbol{Q}_2^{(n+1)}$, we have

$$\boldsymbol{U}_k^{(n)} = \boldsymbol{Q}_1^{(n)} + \begin{pmatrix} \boldsymbol{O} \\ \widetilde{\boldsymbol{r}}_{k-1}^{(n+1)} \end{pmatrix}, \qquad n,\ k \in \mathbb{N}, \tag{5}$$

where

$$\widetilde{\boldsymbol{r}}_k^{(n)} = n\mu\left(0, r_{k,0}^{(n)}, r_{k,1}^{(n)}, \ldots, r_{k,c-1}^{(n)}\right), \qquad n,\ k \in \mathbb{N}.$$

We also have

$$\boldsymbol{R}_k^{(n)} = \boldsymbol{Q}_0^{(n-1)}\left(-\boldsymbol{U}_k^{(n)}\right)^{-1}, \qquad n,\ k \in \mathbb{N},$$

which is equivalent to

$$\boldsymbol{R}_k^{(n)}\boldsymbol{U}_k^{(n)} = -\boldsymbol{Q}_0^{(n-1)}, \qquad n,\ k \in \mathbb{N}. \tag{6}$$

Because the first c rows of both sides of (6) are zero vectors, (6) is equivalent to

$$(x_0, x_1, \ldots, x_c)\boldsymbol{U}_k^{(n)} = (0, 0, \ldots, 0, -\lambda), \tag{7}$$

where $(x_0, x_1, \ldots, x_c)$ is used instead of $\boldsymbol{r}_k^{(n)}$ for convenience. From (5), we can solve (7) efficiently. Indeed, we rewrite (7) as the following system of equations.

$$b_0^{(n)}x_0 + \nu x_1 = 0, \quad i = 0, \tag{8}$$

$$\lambda x_{i-1} + b_i^{(n)}x_i + (i+1)\nu x_{i+1} + (n+1)\mu r_{k-1,i-1}^{(n+1)}x_c = 0, \quad i = 1, 2, \ldots, c-1, \tag{9}$$

$$\lambda x_{c-1} + \left(b_c^{(n)} + (n+1)\mu r_{k-1,c-1}^{(n+1)}\right)x_c = -\lambda, \quad i = c. \tag{10}$$

We assume that x_i $(i = 0, 1, \ldots, c)$ can be expressed in terms of x_c as

$$x_i = \alpha_i + \beta_i x_c, \qquad i = 0, 1, \ldots, c. \tag{11}$$

Substituting (11) into (9) and (10) yields

$$\lambda\alpha_{i-1} + b_i^{(n)}\alpha_i + (i+1)\nu\alpha_{i+1} = 0, \tag{12}$$

$$\lambda\beta_{i-1} + b_i^{(n)}\beta_i + (i+1)\nu\beta_{i+1} + (n+1)\mu r_{k-1,i-1}^{(n+1)} = 0, \tag{13}$$

for $i = c-1, c-2, \dots, 1$. It follows from (10) and (11) with $i = c$ that

$$\alpha_c = 0, \quad \beta_c = 1, \quad \alpha_{c-1} = -1, \quad \beta_{c-1} = -\frac{b_c^{(n)} + (n+1)\mu r_{k-1,c-1}^{(n+1)}}{\lambda}. \tag{14}$$

Substituting (11) into (8) yields

$$b_0^{(n)}(\alpha_0 + \beta_0 x_c) + \nu(\alpha_1 + \beta_1 x_c) = 0. \tag{15}$$

Theorem 1 can be proved by using equations (11) to (15).

Corollary 1 *The solution of the system of equations:*

$$\boldsymbol{x}_0(\boldsymbol{Q}_1^{(0)} + \boldsymbol{R}^{(1)}\boldsymbol{Q}_2^{(1)}) = \boldsymbol{0}, \qquad \boldsymbol{x}_0\boldsymbol{e} = 1, \tag{16}$$

is given by

$$\begin{aligned} x_{c,0} &= \frac{1}{\beta_0 + \beta_1 + \cdots + \beta_{c-1} + 1}, \\ x_{i,0} &= \beta_i x_{c,0}, \qquad i = 0, 1, \dots, c-1, \end{aligned}$$

where $\boldsymbol{x}_0 = (x_{0,0}, x_{1,0}, \dots, x_{c,0})$.

Proof We observe that the system of linear equations:

$$\boldsymbol{x}_0(\boldsymbol{Q}_1^{(0)} + \boldsymbol{R}^{(1)}\boldsymbol{Q}_2^{(1)}) = \boldsymbol{0},$$

expresses a special case of (7) with $n = 0$ and the λ in the right hand side being equal to 0. Note that for this case, $\alpha_{c-1} = 0$ and thus $\alpha_i = 0$ $(i = 0, 1, \dots, c)$. Therefore, from $x_{i,0} = \beta_i x_{c,0}$ and $\boldsymbol{x}_0\boldsymbol{e} = 1$, Corollary 1 is proved.

Because $\{\alpha_i, \beta_i; i = 0, 1, \dots, c\}$ grow fast and the order of α_i and β_i is the same, we confirm that the computation of x_i $(i = 0, 1, \dots, c-1)$ by (11) is numerically unstable. Instead of using (11), we use the following theorem to determine x_i $(i = 0, 1, \dots, c-1)$ provided that x_c is given.

Theorem 2 *If x_c is given, then x_i $(i = 0, 1, \dots, c-1)$ can be determined by*

$$x_i = \frac{(i+1)\nu x_{i+1} + D_i}{B_i}, \qquad i = 0, 1, \dots, c-1, \tag{17}$$

where the sequences $\{B_i, D_i; i = 0, 1, \dots, c-1\}$ are recursively defined by

$$\begin{aligned} B_0 &= \lambda + n\mu, \qquad D_0 = 0, \\ B_i &= (\lambda + i\nu + n\mu) - \frac{\lambda i \nu}{B_{i-1}}, \qquad D_i = (n+1)\mu r_{k-1,i-1}^{(n+1)} x_c + \frac{\lambda D_{i-1}}{B_{i-1}}. \end{aligned} \tag{18}$$

Furthermore, we have

$$B_i > \lambda, \qquad D_i > 0, \qquad i = 0, 1, \ldots, c-1. \tag{19}$$

Proof Equation (17) is easily proved using mathematical induction. We show (19) also by mathematical induction. We confirm that (19) is true for $i = 0$. Assuming that (19) is true for all $i = 0, 1, \ldots, m$, where $m = 0, 1, \ldots, c-2$, we prove that (19) is also true for $i = m+1$. Indeed, it follows from (18) that

$$\begin{aligned} B_{m+1} &= \lambda + (m+1)\nu + n\mu - \frac{\lambda(m+1)\nu}{B_m} \\ &> \lambda + (m+1)\nu + n\mu - \frac{\lambda(m+1)\nu}{\lambda} \\ &= \lambda + n\mu > \lambda, \end{aligned}$$

where $B_m > \lambda$ is used in the first inequality. It follows from (18) and

$$r_{k-1,i-1}^{(n+1)} > 0, \qquad x_c > 0, \qquad B_i > 0,$$

that $D_i > 0$ $(i = 0, 1, \ldots, c-1)$.

Remark 3 According to Falin and Templeton [14], recursive formulae (17) and (18) have been used in analyses of not only retrial queues but also numerical solutions of boundary value problems of second-order differential equations. However, to the best of our knowledge, the inequalities in (19) have not been rigorously proven yet. For example, Artalejo and Pozo [4] use a similar procedure as in Theorem 2, where $B_i, D_i > 0$ $(i = 0, 1, \ldots, c-1)$ is claimed without a proof.

Definition 2 Let r_n denote a function such that

$$r_n(\boldsymbol{x}) = \mathrm{Lr}\left(R_n(\boldsymbol{X}(\boldsymbol{x}))\right), \qquad n \in \mathbb{N},$$

where

$$\boldsymbol{X}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{O} \\ \boldsymbol{x} \end{pmatrix}.$$

In the above, $\boldsymbol{x}$ is a row vector with an appropriate dimension and $\mathrm{Lr}(\boldsymbol{Y})$ denotes the last row of matrix $\boldsymbol{Y}$.

It follows from Theorems 1, 2 and Definition 2 that

$$\boldsymbol{r}^{(n)} = r_n(\boldsymbol{r}^{(n+1)}), \qquad \boldsymbol{r}_k^{(n)} = r_n(\boldsymbol{r}_{k-1}^{(n+1)}) = r_n \circ r_{n+1} \circ \circ \circ r_{n+k-1}(\boldsymbol{0}), \tag{20}$$

for all $n, k \in \mathbb{N}$.

3.2 Algorithm

In this section, first we propose an algorithm to compute the rate matrices. Then, we compute an approximation to the joint stationary distribution, provided that the truncation point is given in advance.

3.2.1 The rate matrices

Recently, Phung-Duc et al. [30] proposed an algorithm to compute an approximation $\widehat{\boldsymbol{R}}^{(n)}$ to $\boldsymbol{R}^{(n)}$. Based on (20), we modify the algorithm in [30] to efficiently compute $\widehat{\boldsymbol{r}}^{(n)}$, which is the last row of $\widehat{\boldsymbol{R}}^{(n)}$. In Algorithm 1, $\{k_l; l \in \mathbb{Z}_+\}$ is a strictly increasing sequence of non-negative integers, and $||\boldsymbol{x}||_\infty$ denotes the infinity norm of vector $\boldsymbol{x}$, whose definition is given by

$$||\boldsymbol{x}||_\infty = \max_j |x_j|,$$

where x_j represents the jth entry of $\boldsymbol{x}$.

Table 1 Computation of $\boldsymbol{r}^{(n)}$.

```
Begin Algorithm 1
Input: $\{\boldsymbol{Q}_0^{(n)}, \boldsymbol{Q}_1^{(n)}, \boldsymbol{Q}_2^{(n)}, k_n; n \in \mathbb{Z}_+, \epsilon\}$.
Output: $\{\widehat{\boldsymbol{r}}^{(n)}\}$.
$l = 1$;
Compute $\boldsymbol{r}_{k_1}^{(n)}$ and $\boldsymbol{r}_{k_0}^{(n)}$ using Theorems 1, 2 and (20).
while $||\boldsymbol{r}_{k_l}^{(n)} - \boldsymbol{r}_{k_{l-1}}^{(n)}||_\infty > \epsilon$ do
    $l := l + 1$;
    Compute $\boldsymbol{r}_{k_l}^{(n)}$ and $\boldsymbol{r}_{k_{l-1}}^{(n)}$ using Theorems 1, 2 and (20).
end
$\widehat{\boldsymbol{r}}^{(n)} := \boldsymbol{r}_{k_l}^{(n)}$;
End Algorithm 1
```

Corollary 2 *The computational complexity of each step in Algorithm 1 and of the boundary equation (16) is $O(c)$.*

Proof This corollary is a direct consequence of the proofs of Theorems 1 and 2.

3.2.2 Stationary distribution

In the general case, no closed form for $\{\boldsymbol{\pi}_n; n \in \mathbb{N}\}$ exists. Therefore, we present an algorithm to compute an approximation $\{\widehat{\boldsymbol{\pi}}_n; n = 0, 1, \ldots, N_0\}$ to the stationary distribution $\{\boldsymbol{\pi}_n; n \in \mathbb{N}\}$, where N_0 is a natural number given in advance. See Table 2 for details of the algorithm, which is modified from Algorithm 3 in [30]. In Table 2, $\boldsymbol{x}_n$ is given by

$$\boldsymbol{x}_n = (x_{0,n}, x_{1,n}, \ldots, x_{c,n}), \qquad n = 0, 1, \ldots, N_0,$$

which corresponds to $\boldsymbol{\pi}_n$. We also use

$$\boldsymbol{x}_{n-1}\widehat{\boldsymbol{R}}^{(n)} = x_{c,n-1}\widehat{\boldsymbol{r}}^{(n)},$$

to simplify the algorithm. A careful choice of N_0 for the M/M/c/c retrial queue will be discussed in Section 3.3.

Table 2 The stationary distribution.

```
Begin Algorithm 2
Input: $\lambda, \mu, \nu, c, \{k_n; n \in \mathbb{N}\}, \epsilon, N_0$.
Output: $\{\widehat{\boldsymbol{\pi}}_n; n = 0, 1, \ldots, N_0\}$.
Compute $\widehat{\boldsymbol{r}}^{(N_0)}$ using Algorithm 1 with $\{k_n\}$ and $\epsilon$.
for $n = 1$ to $N_0 - 1$ do
    $\widehat{\boldsymbol{r}}^{(N_0-n)} = r_{N_0-n}(\widehat{\boldsymbol{r}}^{(N_0-n+1)})$;
end
Compute $\boldsymbol{x}_0$ by Corollary 1.
for $n = 1$ to $N_0$ do
    $\boldsymbol{x}_n = x_{c,n-1}\widehat{\boldsymbol{r}}^{(n)}$;
end
for $n = 0$ to $N_0$ do
    $\widehat{\boldsymbol{\pi}}_n := \frac{\boldsymbol{x}_n}{\sum_{n=0}^{N_0} \boldsymbol{x}_n \boldsymbol{e}}$;
end
End Algorithm 2
```

3.3 Choice of the truncation level N_0

In Algorithm 2, the truncation level N_0 is given in advance. It is desired that N_0 is the level where the tail probability is small enough to be neglected. In other words, we need an N_0 such that

$$\sum_{n=N_0+1}^{\infty} \boldsymbol{\pi}_n \boldsymbol{e} < \epsilon_0.$$

However, since $\boldsymbol{\pi}_n$ is unknown, it is difficult to directly determine such N_0 for a general ergodic M/M/c/c retrial queue.

Recall that an explicit solution for the joint stationary distribution of an M/M/1/1 retrial queue is obtained in [14]. We consider an M/M/1/1 retrial queue with an arrival rate λ/c, a service rate ν and a retrial rate μ. This M/M/1/1 retrial queue is also stable because $\rho = \lambda/(c\nu) < 1$. Let $p_{i,n}$ $(i = 0, 1, n \in \mathbb{Z}_+)$ denote the joint stationary probability that there are i busy server and n customers in the orbit. According to [14], we have the following result.

$$p_{0,n} = \frac{\rho^n}{n!}(1-\rho)^{\frac{\lambda}{c\mu}+1}\left(\frac{\lambda}{c\mu}\right)_n, \qquad p_{1,n} = \frac{\rho^{n+1}}{n!}(1-\rho)^{1+\frac{\lambda}{c\mu}}\left(1+\frac{\lambda}{c\mu}\right)_n,$$

for all $n \in \mathbb{Z}_+$, where $(\varphi)_n$ $(-\infty < \varphi < \infty,\ n \in \mathbb{Z}_+)$ denotes the Pochhammer (see e.g. page 222 in [11]), whose definition is given by

$$(\varphi)_n = \begin{cases} 1, & n = 0, \\ \varphi(\varphi+1)\cdots(\varphi+n-1), & n \in \mathbb{N}. \end{cases}$$

The truncation point is determined by

$$N_0 = \inf\{n \mid \sum_{i=0}^{n}(p_{0,i} + p_{1,i}) > 1 - \epsilon_0\}, \tag{21}$$

for any $\epsilon_0 > 0$. According to numerical results in [29], it seems that the tail probability of the M/M/1/1 retrial queue is greater than that of the M/M/c/c retrial queues ($c = 2, 3$ and 4) with the same traffic intensity. Based on these observations, we further expect that the tail probability of an M/M/c/c ($c \geq 2$) is also smaller than that of the M/M/1/1 retrial queue. In particular,

$$\sum_{n=N_0+1}^{\infty} \boldsymbol{\pi}_n \boldsymbol{e} < \sum_{n=N_0+1}^{\infty} (p_{0,n} + p_{1,n}) < \epsilon_0.$$

This supports the choice of N_0 by (21). The ergodic condition of the M/M/1/1 retrial queue is the same as that of the M/M/c/c retrial queue. Therefore, we can obtain N_0 for any λ and ν satisfying $\lambda < c\nu$, which is equivalent to the ergodic condition $\rho < 1$.

4 Performance Measures and Numerical Examples

In this section, we derive some performance measures and then provide some numerical results.

4.1 Performance measures

Let π_n denote the probability that there are n customers in the orbit in the steady state. We have

$$\pi_n = \sum_{i=0}^{c} \pi_{i,n}, \qquad n \in \mathbb{Z}_+.$$

Let E$[C]$, Var(C), E$[N]$ and B denote the average and the variance of the number of busy servers, the average number of customers in the orbit and the blocking probability, respectively. We have

$$\mathrm{E}[C] = \sum_{n\in\mathbb{Z}_+}\sum_{i=0}^{c}\pi_{i,n}i, \qquad \mathrm{E}[N] = \sum_{n\in\mathbb{Z}_+} n\pi_n, \qquad B = \sum_{n\in\mathbb{Z}_+} \pi_{c,n}.$$

and

$$\mathrm{Var}(C) = \sum_{n=0}^{\infty}\sum_{i=0}^{c}\pi_{i,n}i^2 - \mathrm{E}[C]^2.$$

Remark 4 The blocking probability B is defined as the probability that an arriving primary or retrial customer finds all the servers busy.

Let B_{wor} denote the blocking probability of the M/M/c/c Erlang loss system with arrival rate λ and service rate ν. We have

$$B_{wor} = \frac{\frac{(c\rho)^c}{c!}}{\sum_{i=0}^{c}\frac{(c\rho)^i}{i!}}.$$

Let B_{wr} denote the blocking probability of the M/M/c/c Erlang loss system with arrival rate λ^* and service rate ν, where

$$\lambda^* = \lambda + \mu\mathrm{E}[N].$$

λ^* expresses the average arrival rate of the primary customers and retrial customers. We have

$$B_{wr} = \frac{\frac{(c\rho^*)^c}{c!}}{\sum_{i=0}^{c}\frac{(c\rho^*)^i}{i!}},$$

where

$$\rho^* = \frac{\lambda^*}{c\nu}.$$

Let $\widehat{\pi}_n$, $\widehat{\mathrm{E}}[C]$, $\widehat{\mathrm{E}}[N]$ and $\widehat{B}$ denote the approximations to $\pi_n, \mathrm{E}[C]$, E$[N]$ and B, respectively, i.e.,

$$\widehat{\pi}_n = \sum_{i=0}^{c}\widehat{\pi}_{i,n}, \quad \widehat{\mathrm{E}}[C] = \sum_{n=0}^{N_0}\sum_{i=0}^{c}\widehat{\pi}_{i,n}i, \quad \widehat{\mathrm{E}}[N] = \sum_{j=0}^{N_0} n\widehat{\pi}_n, \quad \widehat{B} = \sum_{n=0}^{N_0}\widehat{\pi}_{c,n},$$

and further let

$$\widehat{\mathrm{Var}}(C) = \sum_{n=0}^{N_0}\sum_{i=0}^{c}\widehat{\pi}_{i,n}i^2 - \widehat{\mathrm{E}}[C]^2,$$

denote an approximation to $\mathrm{Var}(C)$. According to Falin and Templeton [14], we have some explicit formulae as follows:

$$\mathrm{E}[C] = \lambda, \qquad \mathrm{E}[N] = \frac{1+\mu}{\mu} \frac{\lambda - \mathrm{Var}(C)}{c - \lambda}, \tag{22}$$

provided that $\nu = 1$. We define the absolute errors e_C and e_N by

$$e_C = \left| \widehat{\mathrm{E}}[C] - \lambda \right|, \qquad e_N = \left| \widehat{\mathrm{E}}[N] - \frac{1+\mu}{\mu} \frac{\lambda - \widehat{\mathrm{Var}}(C))}{c - \lambda} \right|,$$

provided that $\nu = 1$, in order to evaluate the accuracy of our algorithm.

4.2 Numerical examples

In this section, we present some numerical examples to show the efficiency of our algorithm and to evaluate the performance of the M/M/c/c retrial queue.

4.2.1 Influence of ϵ in Algorithm 1

We consider an example where $c = 10$, $\mu = 1$ and $\nu = 1$ to investigate the influence of ϵ on the stationary distribution obtained by Algorithm 2. Fig. 2 shows the absolute error e_C against the truncation point N_0 in Algorithm 2. In Fig. 2, the four pairs of curves from the left to the right correspond to the cases of $\rho = 0.5, 0.7, 0.9$ and 0.95, respectively. In all the curves, e_C decreases with N_0. In each pair (for example, the pair of $\rho = 0.9$), the left and the right curves correspond to the case where $\boldsymbol{r}_{N_0+1} = \mathbf{0}$, and the case where $\boldsymbol{r}_{N_0}$ is computed using Algorithm 1 with $k_l = 2^l - 1$ and $\epsilon = 10^{-7}$, respectively. We observe that the absolute error of the latter is smaller than that of the former, although the difference between both curves is small. This implies that the impact of ϵ on the stationary distribution is small and that the accuracy brought by Algorithm 2 is insensitive to ϵ.

The truncation points computed by the procedure in Section 3.3 with $\epsilon_0 = 10^{-7}$ are 24, 49, 177 and 368 for the cases of $\rho = 0.5, 0.7, 0.9$ and 0.95, respectively. Fig. 2 shows that at these truncation points, even for the cases where $\boldsymbol{r}_{N_0+1} = \mathbf{0}$, e_C is in the order of 10^{-6}. This implies that for the N_0 computed by the procedure in Section 3.3, the simple method where $\boldsymbol{r}_{N_0+1} = \mathbf{0}$ has a sufficient accuracy. Therefore, in Section 4.2, instead of computing $\boldsymbol{r}_{N_0}$ with high accuracy, we choose the N_0 as presented in Section 3.3 with $\epsilon_0 = 10^{-7}$ and use $\boldsymbol{r}_{N_0+1} = \mathbf{0}$. The validation of the N_0 will be presented in details in Section 4.2.6

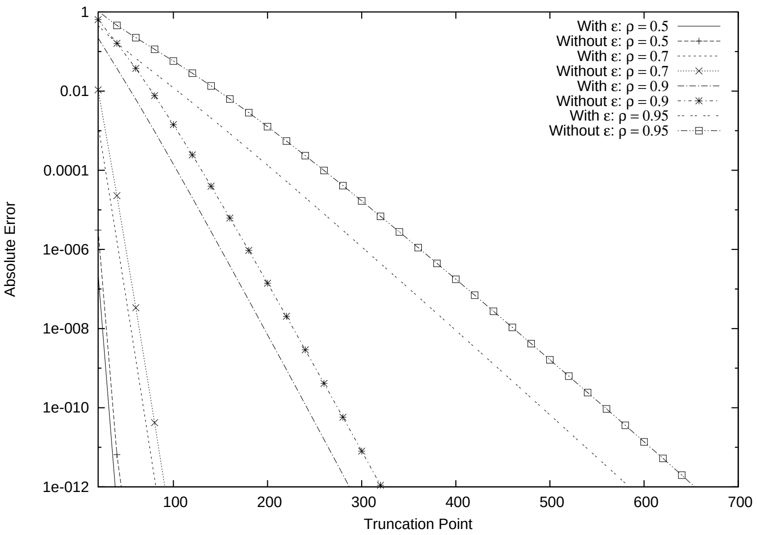

Fig. 2 Absolute Error e_C vs. Truncation Point.

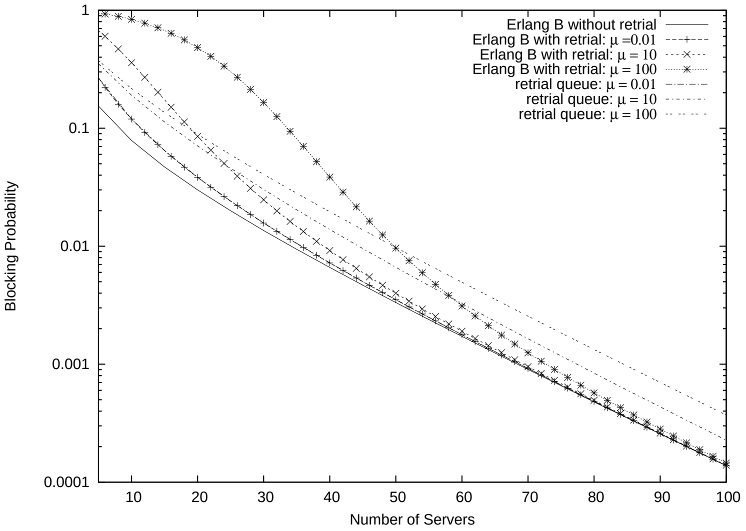

Fig. 3 Blocking Probability vs. the Number of Servers ($\rho = 0.7$).

4.2.2 Optimal number of servers

In the following, we use $\epsilon_0 = 10^{-7}$ to determine the truncation point N_0 by the procedure presented in Section 3.3 and $\nu = 1$. Now we compare three types of blocking probabilities B_{wor}, B and B_{wr}.

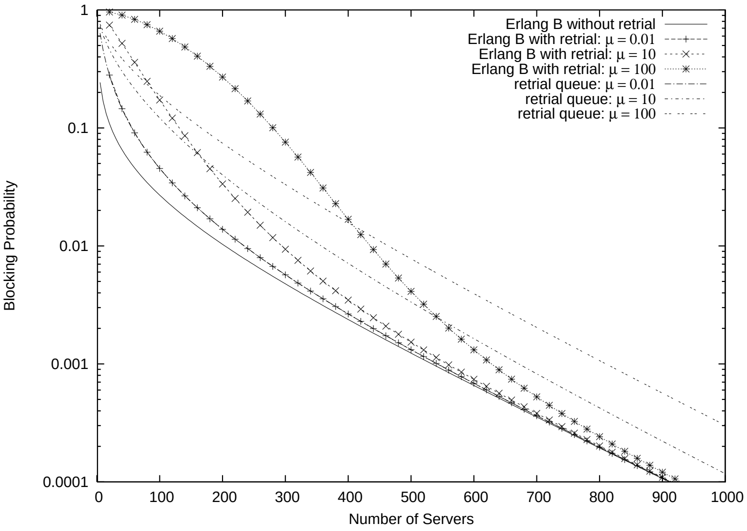

Fig. 4 Blocking Probability vs. the Number of Servers ($\rho = 0.9$).

Figs. 3 and 4 demonstrate B_{wor}, B and B_{wr} against the number of servers for two cases: $\rho = 0.7$ and 0.9. In these figures, the blocking probability without retrials B_{wor} is the smallest. We observe that under the same number of servers, B and B_{wr} increase with μ as expected.

We compare the curves of B_{wr} and B with the same μ. When μ is small, e.g., $\mu = 0.01$, both curves are almost the same. The difference between them increases with μ. In the case where μ is large, e.g. $\mu = 100$, we observe that there exists some c_0 such that $B \leq B_{wr}$ and $B \geq B_{wr}$ provided that $c \leq c_0$ and $c \geq c_0$, respectively. The reason for this can be explained as follows.

In case of a small c, the probability that all the servers are busy is large. Thus, in the M/M/c/c retrial queue, blocking is easily observed by retrial customers repeatedly. On the other hand, in M/M/c/c loss model, primary customers and retrial customers observe the same congested situation. These reasons explain why $B_{wr} > B$. In the case of a large c, the probability that all the servers are busy is small. Therefore, B_{wr} is small due to the fact that in the Erlang loss model, customers are assumed to arrive at the servers randomly. In the M/M/c/c retrial queue, if a customer is blocked, the customer retries immediately. Thus, the arrival process of retrial customers is likely to have a bursty nature which results in a high blocking probability B.

We consider an optimal design problem using B_{wor}, B_{wr} and B. Let c_ϵ denote the minimum number of servers such that the blocking probability is less than or equal to ϵ. In the case where $\mu = 100$, we find from the curves of B_{wor} and B_{wr} in Fig. 3 that

$c_{0.1} \approx 8$ and 33, respectively, while the curve of B shows that $c_{0.1} = 18$. These results show that B_{wor} underestimates and B_{wr} overestimates the optimal number of servers.

The difference between the curves of B and B_{wr} for the case of $\rho = 0.9$ is even larger than that for the case of $\rho = 0.7$. We find from Fig. 4 that $c_{0.1}$'s for the curves of B and B_{wr} are roughly 180 and 280, respectively. As for the case of $c_{0.001}$, the answers given by B and B_{wr} are about 800 and 620, respectively. Therefore, B_{wr} overestimates $c_{0.1}$ and underestimates $c_{0.001}$.

From the observations on Figs. 3 and 4, our conclusion is as follows. For the case of a small retrial rate, e.g. $\mu < 1$, the Erlang B formula gives a good estimation of the optimal number of servers. However, when the retrial rate is large, e.g. $\mu > 1$, the estimation by the Erlang B formula has a large error. Therefore, in applications such as call centers, cellular mobile networks, etc., where the retrial interval is short, a retrial queueing model should be used instead of a conventional Erlang loss model.

4.2.3 Influence of the retrial rate

Figs. 5 and 6 show B and B_{wr} against μ for the cases of $\rho = 0.9$ and 0.95, respectively. In these figures, the curves for $c = 10, 50, 100, 500$ and 1000 are plotted.

First, we compare B and B_{wr} in a wide range of μ. We observe in all the cases that there exist some μ_0 and μ_1 ($\mu_0 \leq \mu_1$) such that $B \approx B_{wr}$, $B \geq B_{wr}$ and $B \leq B_{wr}$ provided that $\mu \leq \mu_0$, $\mu_0 \leq \mu \leq \mu_1$ and $\mu \geq \mu_1$, respectively. In particular, Figs. 5 and 6 show that when μ is large B is insensitive to μ while B_{wr} is sensitive to μ, and that both B and B_{wr} are insensitive to μ when μ is small.

Second, we consider the characteristics of B for M/M/c/c retrial queues. For example, Fig. 5 shows that B is insensitive to μ when $\mu < 1$. This suggests that we can use the result of B for $\mu = 1$ to obtain an approximation of B for a smaller μ. This saves on computational cost because a small μ causes a large N_0. We further observe that B comes closer to the probability that a customer has to wait in the conventional M/M/c queue as shown in [14].

4.2.4 Average number of customers in the orbit

Fig. 7 demonstrates the average number of customers in the orbit against the number of servers for the case of $\rho = 0.95$. In Fig. 7, the curves of the cases, where $\mu = 0.01, 0.1, 1, 10$ and 100 are plotted. We observe in these curves that the average number of retrial customers decreases with the number of servers. This is due to the collaboration among the servers. It should be noted that the vertical axis of the graphs is in log-scale. In this scale we observe that the average number of customers in the orbit is asymptotically linear with the number of servers. As a result, we can predict E[N] for the case of a large number of servers using the E[N] obtained for the case of a

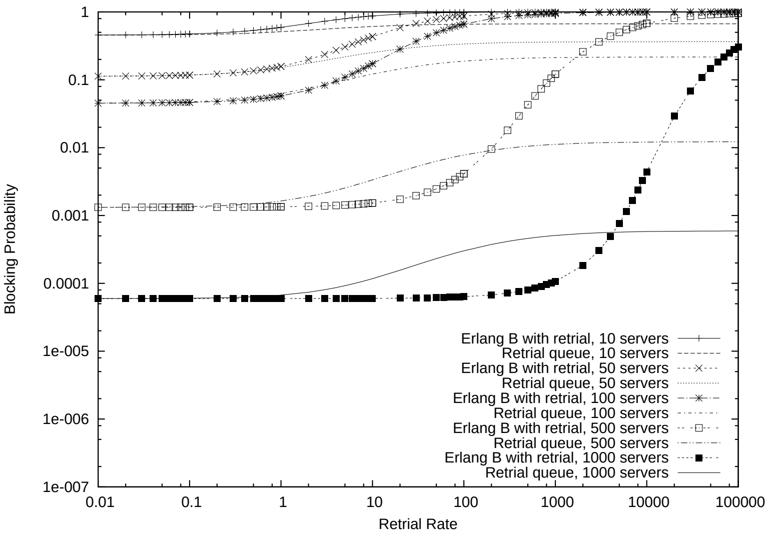

Fig. 5 Blocking Probability vs. Retrial Rate ($\rho = 0.9$).

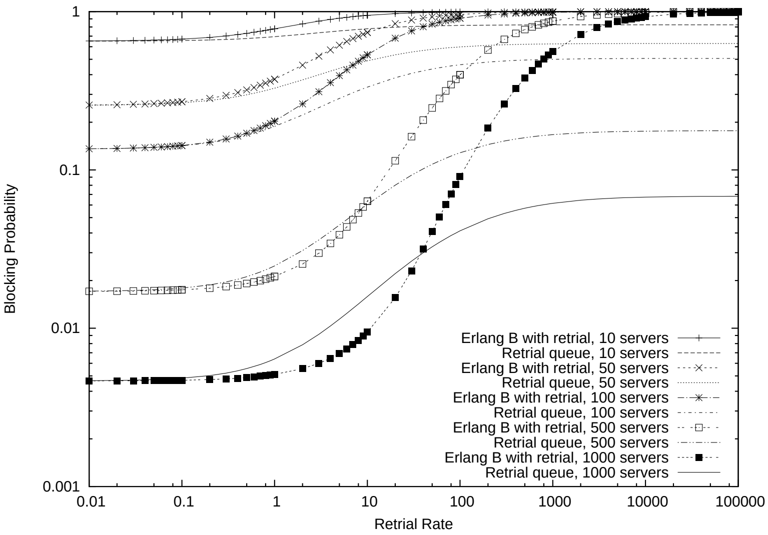

Fig. 6 Blocking Probability vs. Retrial Rate ($\rho = 0.95$).

relatively small number of servers due to the fact that a line is completely characterized by a point and a slope.

We investigate the influence of μ on the average number of retrial customers. We observe in Fig. 8 that the average number of retrial customers decreases with the retrial rate and is asymptotic to the average number of waiting customers in the corresponding

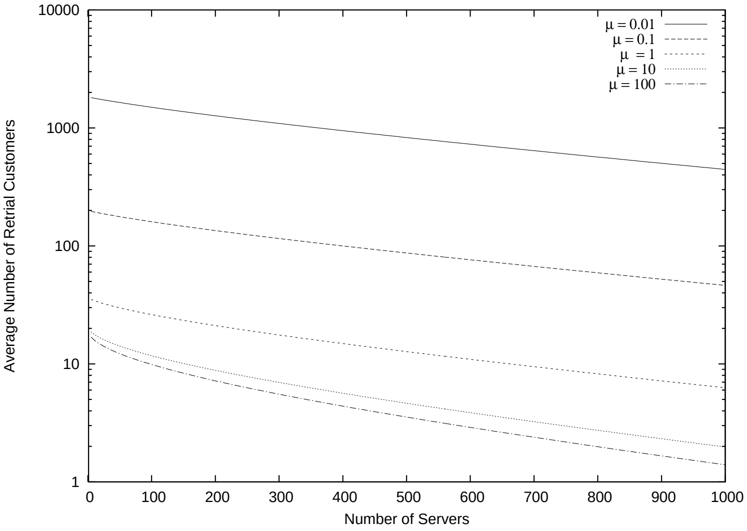

Fig. 7 Average Number of Retrial Customers vs. the Number of Servers ($\rho = 0.95$).

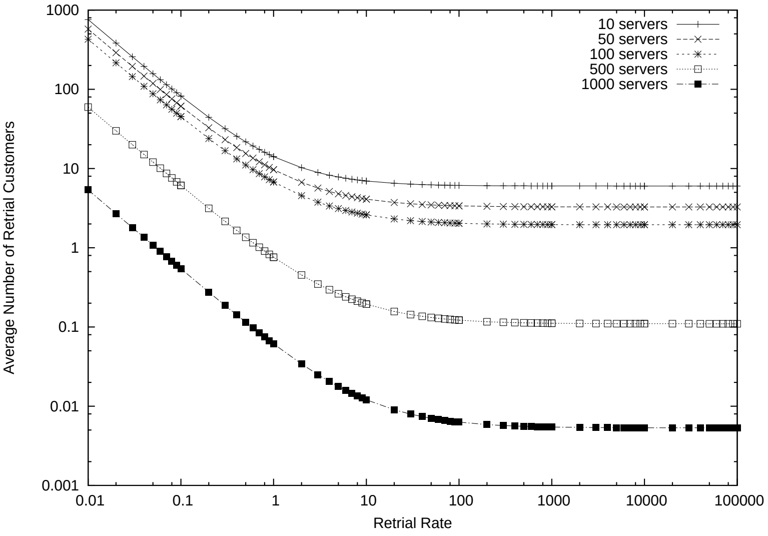

Fig. 8 Average Number of Retrial Customers vs. Retrial Rate ($\rho = 0.9$).

conventional M/M/c queue without retrial. Note that Fig. 8 is plotted in log-scale. In this scale, the curves show a linear tendency when $\mu < 1$, which agrees with the fact that E[N] is proportional to μ^{-1} as $\mu \to \infty$ as shown in [14]. Based on this fact, we can use the E[N] obtained for the case of $\mu = 1$ in order to obtain an approximation to the E[N] for the case of a smaller μ.

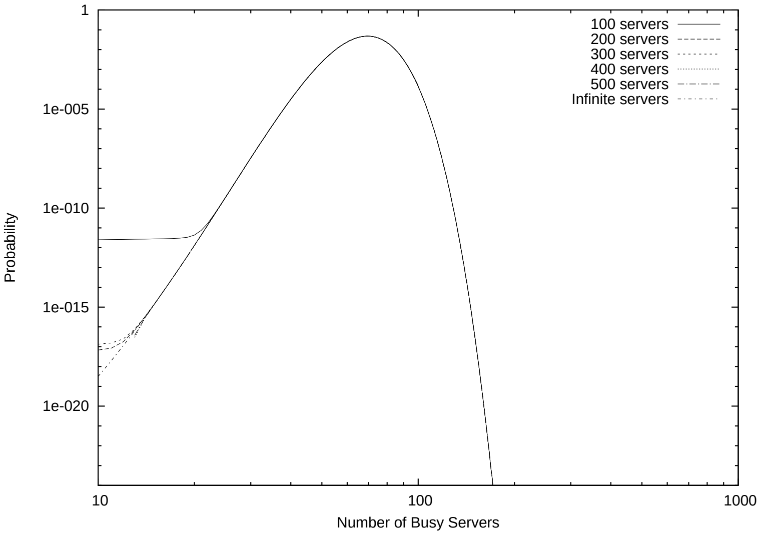

Fig. 9 Asymptotic of the Distribution of the Number of Busy Servers

4.2.5 Asymptotic behavior

Fig. 9 shows the asymptotic behavior of the distribution of the number of busy servers when the arrival rate and the service rate are kept constant, $\lambda = 70$ and $\nu = 1$. We observe that the stationary distribution of the number of busy servers in M/M/c/c retrial queues is asymptotically the same as the stationary distribution of the number of busy servers in an M/M/∞ queue with the same arrival and service rates. The reason for this is that customers are not blocked when the number of servers is large enough. Therefore, the retrial queue behaves like the M/M/∞ queue. This suggests that if the number of servers is large enough, we can use an M/M/∞ queue instead of a multiserver retrial queue.

4.2.6 Validation of the truncation point N_0

In this section, we show the validation of the determination of N_0 presented in 3.3. Fig. 10 represents the absolute error e_N against the traffic intensity for the cases of $\mu = 1$ and $c = 10, 100$ and 500. The figure shows that the absolute error is small when the traffic intensity is small, while they become larger as ρ approaches to 1. However, under such a heavy traffic condition, the number of retrial customers is also large, thus, the relative errors are small.

Fig. 11 shows $\widehat{\mathrm{E}}[C]$ against the number of servers for the cases of $\rho = 0.7, 0.9$ and 0.95. We observe that in all the cases, $\widehat{\mathrm{E}}[C]$ does not depend on μ. We also confirm

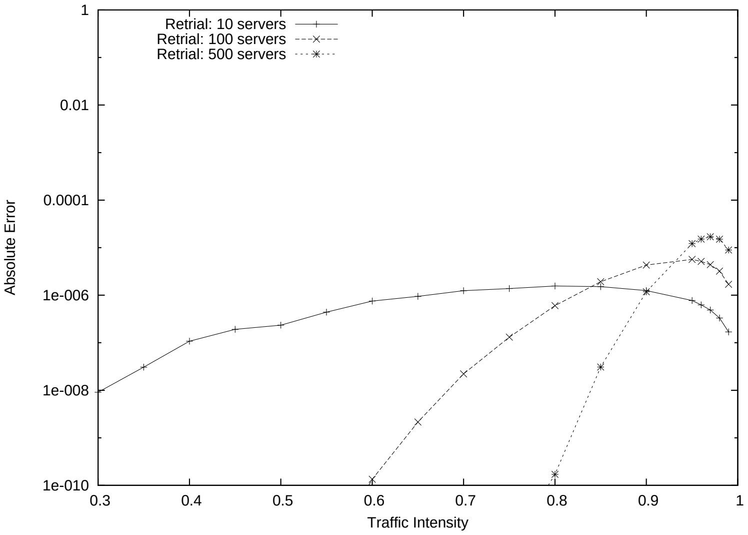

Fig. 10 Absolute Error e_N vs. Traffic Intensity.

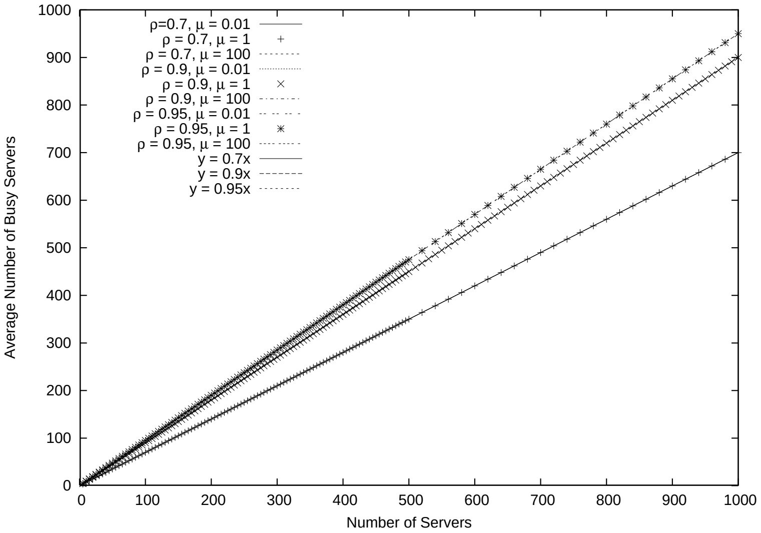

Fig. 11 Average Number of Busy Servers vs. the Number of Servers.

that the curves of $\widehat{\mathrm{E}}[C]$ conform with the line $y = \rho c$, where ρ is kept constant. These results agree with (22).

5 Conclusion

In this paper, we have presented an algorithm to compute the stationary distribution of the M/M/c/c retrial queue. The algorithm is based on a matrix continued fraction representation of the rate matrices of the level-dependent QBD underlying the queue. One of the most notable features of the algorithm is that it does not need to compute inverse matrices. The computational complexity of the algorithm is only $O(c)$ instead of $O(c^3)$ as in conventional matrix analytic methods [30]. Furthermore, we have shown that the algorithm manipulates positive numbers and thus it is numerically stable. This enables us to analyze M/M/c/c retrial queues with large c. We believe that the algorithm can be applied to some variant models such as state-dependent multiserver retrial queues with abandonments.

Acknowledgements This research is partially supported by Japan Society for the Promotion of Science, Grant-in-Aid for JSPS Fellows (No. 22·470).

References

1. S. Aguir, F. Karaesmen, O. Z. Aksin and F. Chauvet (2004). The impact of retrials on call center performance. *OR Spectrum, 26*, 353–376.
2. A. S. Alfa and W. Li (2002). PCS networks with correlated arrival process and retrial phenomenon. *IEEE Transactions on Wireless Communications, 1*, 630–637.
3. V. V. Anisimov and J. R. Artalejo (2002). Approximation of multiserver retrial queues by means of generalized truncated models. *TOP: An Official Journal of the Spanish Society of Statistics and Operations Research, 10*, 51–66.
4. J. R. Artalejo and M. Pozo (2002). Numerical calculation of the stationary distribution of the main multiserver retrial queue. *Annals of Operations Research, 116*, 41–56.
5. J. R. Artalejo and A. Gomez-Corral (2008). *Retrial queueing systems: a computational approach*. Berlin Heidelberg: Springer-Verlag.
6. J. R. Artalejo and V. Pla (2009). On the impact of customer balking, impatience and retrials in telecommunication systems. *Computer and Mathematics with Applications, 57*, 217–229.
7. J. R. Artalejo and M. J. Lopez-Herrero (2010). Cellular mobile networks with repeated calls operating in random environment. *Computers & Operations Research, 37*, 1158-1166.
8. H. Baumann and W. Sandmann (2010). Numerical solution of level dependent quasi-birth-and-death processes. *Procedia Computer Science, 1*, 1555–1563.
9. L. Breuer, A. Dudin and V. Klimenok (2002). A retrial BMAP/PH/N system. *Queueing Systems, 40*, 433–457.
10. L. Bright and P. G. Taylor (1995). Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models, 11*, 497–525.
11. A. Cuyt, V. B. Petersen, B. Verdonk, H. Waadeland and W. B. Jones (2008). *Handbook of continued fractions for special functions.* Springer Science+Business Media B.V.
12. B. D. Choi, Y. Chang and B. Kim (1999). MAP1, MAP2/M/c retrial queue with guard channels and its application to cellular networks. *TOP: An Official Journal of the Spanish Society of Statistics and Operations Research, 7*, 231-248.

13. M. J. Domenech-Benlloch, J. M. Gimenez-Guzman, V. Pla, J. Martinez-Bauset and V. Casares-Giner (2008). Generalized truncated methods for an efficient solution of retrial systems. *Mathematical Problems in Engineering*, doi:10.1155/2008/183089.
14. G. I. Falin and J. G. C. Templeton (1997). *Retrial queues.* Chapman & Hall.
15. W. Fair (1971). Noncommutative continued fractions. *SIAM Journal of Mathematical Analysis, 2*, 226–232.
16. W. Fair (1972). A convergence theorem for noncommunitative continued fractions. *Journal of Approximation Theory, 5*, 74–76.
17. A. Gomez-Corral (2006). A bibliographical guide to the analysis of retrial queues through matrix analytic techniques. *Annals of Operations Research, 141*, 163–191.
18. T. Hanschke (1987). Explicit formulas for the characteristics of the M/M/2/2 queue with repeated attempts. *Journal of Applied Probability, 24*, 486–494.
19. T. Hanschke (1999). A matrix continued fraction algorithm for the multiserver repeated order queue. *Mathematical and Computer Modelling, 30*, 159–170.
20. V. Klimenok and A. Dudin (2006). Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems, 54*, 245–259.
21. G. Koole and A. Mandelbaum (2002). Queueing models of call centers: an introduction. *Annals of Operations Research, 113*, 41–59.
22. N. Gans, G. Koole and A. Mandelbaum (2003). Telephone call centers: tutorial, review, and research prospects. *Manufacturing & Service Operations Management, 5*, 79–141.
23. B. Liu and Y.Q. Zhao (2010). Analyzing retrial queues by censoring. *Queueing Systems, 64*, 203–225.
24. M. A. Marsan, G. de Carolis, E. Leonardi, R. Lo Cigno and M. Meo (2001). Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials. *IEEE Journal on Selected Areas in Communications, 19*, 332–346.
25. M. F. Neuts and B. M. Rao (1990). Numerical investigation of a multiserver retrial model. *Queueing Systems, 7*, 169–190.
26. S. T. Peng and A. Hessel (1975). Convergence of noncommutative continued fraction. *SIAM Journal of Mathematical Analysis, 6*, 724–727.
27. T. Phung-Duc, H. Masuyama, S. Kasahara and Y. Takahashi (2009). Performance analysis of optical burst switched networks with limited-range wavelength conversion, retransmission and burst segmentation. *Journal of the Operations Research Society of Japan, 52*, 58–74.
28. T. Phung-Duc, H. Masuyama, S. Kasahara and Y. Takahashi (2009). M/M/3/3 and M/M/4/4 retrial queues. *Journal of Industrial and Management Optimization, 5*, 431–451.
29. T. Phung-Duc, H. Masuyama, S. Kasahara and Y. Takahashi (2010). State-dependent M/M/c/$c+r$ retrial queues with Bernoulli abandonment. *Journal of Industrial and Management Optimization, 6*, 517–540.
30. T. Phung-Duc, H. Masuyama, S. Kasahara and Y. Takahashi (2010). A simple algorithm for the rate matrices of level-dependent QBD processes. In *Proceedings of the 5th International Conference on Queueing Theory and Network Applications.*
31. P. Tran-Gia and M. Mandjes (1997). Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE Journal on Selected Areas in Communications, 15*, 1406–1414.
32. S. N. Stepanov (1999). Markov models with retrials: the calculation of stationary performance measures based on concept of truncation. *Mathematical and Computer Modelling, 30*, 207–228.