Applied Mathematics and Computation 235 (2014) 124–136

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Maximum queue lengths during a fixed time interval in the $M/M/c$ retrial queue ☆

A. Gómez-Corral [a,*], M. López García [b]

[a] Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, 28040 Madrid, Spain
[b] Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds LS2 9JT, United Kingdom

ARTICLE INFO

Keywords:
Absorbing Markov chain
Eigenvalues/eigenvectors
Maximum queue length
Retrial queue
Splitting method

ABSTRACT

We are concerned with the problem of characterizing the distribution of the maximum number $Z(t_0)$ of customers during a fixed time interval $[0, t_0]$ in the $M/M/c$ retrial queue, which is shown to have a matrix exponential form. We present a simple condition on the service and retrial rates for the matrix exponential solution to be explicit or algorithmically tractable. Our methodology is based on splitting methods and the use of eigenvalues and eigenvectors. A particularly appealing feature of our solution is that it allows us to obtain global error control. Specifically, we derive an approximating solution $p(x;t_0) \equiv p(x;t_0;\varepsilon)$ verifying $|P(Z(t_0) \leqslant x|X(0) = (i,j)) - p(x;t_0)| < \varepsilon$ uniformly in $x \geqslant i + j$, for any $\varepsilon > 0$ and initial numbers i of busy servers and j of customers in orbit.

1. Introduction

In this paper we consider maximum queue lengths in the $M/M/c$ retrial queue. The $M/M/c$ retrial queue is the main multiserver retrial queue (see [11, Chapter 2]) in which primary customers arrive according to a Poisson process of rate λ, the service facility consists of c identical servers, and service times are exponentially distributed with parameter ν. If a primary customer finds some server free, he instantly occupies one server and leaves the system after service. Any customer who finds all servers busy upon arrival is obliged to leave the service area, but he repeats his demand after an exponential time with parameter μ; i.e., inter-retrial times of each customer are assumed to be independent and exponentially distributed with intensity μ. We assume that inter-arrival times, service times and inter-retrial times are mutually independent.

The system state at time t can be described by means of a bivariate process $\mathcal{X} = \{X(t) = (C(t), N(t)) : t \geqslant 0\}$, where $C(t)$ is the number of busy servers and $N(t)$ is the number of customers in orbit, that is, customers repeating their demand. Under the above distributional assumptions, the process $\mathcal{X}$ is a regular continuous-time Markov chain (CTMC) with the lattice semistrip $\mathcal{S} = \{0, 1, \ldots, c\} \times \mathbb{N}_0$ as the state space. Its non-null infinitesimal transition rates $q_{(i,j),(i',j')}$ are specified as follows:

☆ Financial support for this work was provided by the Government of Spain (Ministry of Economy and Competitiveness) and the European Commission through the project MTM-2011-23864, and the Grant BES-2009-018747.

* Corresponding author.
E-mail addresses: antonio_gomez@mat.ucm.es (A. Gómez-Corral), M.LopezGarcia@leeds.ac.uk (M. López García).
URLs: http://www.mat.ucm.es/~mcqt/qmg/gomezc.html (A. Gómez-Corral), https://www1.maths.leeds.ac.uk/~lopezgarcia/index.html (M. López García).

http://dx.doi.org/10.1016/j.amc.2014.02.074
0096-3003/

(a) For $0 \leqslant i \leqslant c-1$,

$$q_{(i,j),(i',j')} = \begin{cases} \lambda, & \text{if } (i',j') = (i+1,j), \\ i\nu, & \text{if } (i',j') = (i-1,j), \\ j\mu, & \text{if } (i',j') = (i+1,j-1), \end{cases} \tag{1}$$

and $q_{(i,j)} = -q_{(i,j),(i,j)} = \lambda + i\nu + j\mu$.

(b) For $i = c$,

$$q_{(c,j),(i',j')} = \begin{cases} \lambda, & \text{if } (i',j') = (c,j+1), \\ c\nu, & \text{if } (i',j') = (c-1,j), \end{cases} \tag{2}$$

and $q_{(c,j)} = -q_{(c,j),(c,j)} = \lambda + c\nu$.

If we express $\mathcal{S}$ in terms of *levels* $\mathcal{S} = \cup_{j=0}^{\infty} l(j)$ with $l(j) = \{(i, j-i) : 0 \leqslant i \leqslant \min\{j,c\}\}$ for $j \geqslant 0$, then the infinitesimal generator $\mathbf{Q} = (q_{(i,j),(i',j')})$ of the process $\mathcal{X}$ has the structured form

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,1} & & & \\ \mathbf{A}_{1,0} & \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & & \\ & \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \mathbf{A}_{2,3} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{3}$$

where $\mathbf{A}_{j,j'}$ contains transition rates related to jumps of $\mathcal{X}$ from states of $l(j)$ to states of the level $l(j')$, for $j' \in \{j-1, j, j+1\}$, and diagonal elements of $\mathbf{A}_{j,j}$ are given by $-q_{(i,j-i)}$, for $0 \leqslant i \leqslant \min\{j,c\}$. Specifications for $\mathbf{A}_{j,j'}$ are readily derived from (1) and (2); see Appendix A.

The main analytical difficulties and the most interesting properties of the $M/M/c$ retrial queue are connected with the level dependence exhibited by the infinitesimal generator $\mathbf{Q}$ in (3). To show the nature of these difficulties in more detail, we may consider the simplest problem, that is, the calculation of the stationary distribution $\{P_{i,j} : (i,j) \in \mathcal{S}\}$ of the process $\mathcal{X}$ under the assumption that the traffic load $\rho = (c\nu)^{-1}\lambda$ is less than one. It is worth mentioning that, for $c \leqslant 2$, the partial sequences $\{P_{i,j} : j \in \mathbb{N}_0\}$ with $0 \leqslant i \leqslant c$ satisfy sets of equations of *birth-and-death* type and, consequently, explicit expressions for the stationary probabilities $P_{i,j}$ are recursively derived. The consideration of more than two servers complicates the transitions among states, which implies that the underlying structure of birth-and-death type is not preserved. The particular case $c = 3$ is treated in [13], where the problem is reduced to finding the probabilities $P_{0,0}$ and $P_{0,1}$, which can be recursively computed in terms of a limit condition. The papers by Phung-Duc et al. [23,24] overcome this technical condition and investigate in more detail the stationary probabilities $P_{i,j}$. Based on the Kolmogorov equations some theoretical approaches provide solutions in terms of contour integrals [8] or as limit of extended continued fractions [22]; see [26] for a matrix application of the continued fraction approach to multiserver retrial queues. However, from a practical point of view, the stationary probabilities $P_{i,j}$ cannot be expressed in a tractable form and do not lead to a direct recursive computation when $c > 3$. This drawback motivates the implementation of numerically tractable approximations, such as approximations based on truncated models [11,28,31] and generalized truncated models [2,10,21], the RTA (*retrials see time averages*) approximation due to Greenberg and Wolff [15,32], the Fredericks and Reisner approximation [12], and approximations by interpolation [5,27]. The book by Artalejo and Gómez-Corral [5] presents a comparative review of these approximations for the $M/M/c$ retrial queue, as well as of other performance measures. In the recent monograph by Dayar [9] and the paper by Bright and Taylor [7] the reader can find details on computational issues for an appropriate numerical evaluation of stationary distributions in the more general setting of Markov chains and level-dependent QBD processes. A recent development of algorithmic tools in level-dependent QBD processes can be found in [25].

In this paper, we aim to study the maximum queue length $Z(t_0)$ during a fixed time interval $[0, t_0]$ in the $M/M/c$ retrial queue defined by (1) and (2). One way of analyzing the maximum size distribution is to record maximum values during a busy period, instead of a predetermined time interval. Let us recall that, in the $M/M/c$ retrial queue, a busy period is defined as the period $[0, T]$ that, starting with the arrival of a customer who finds the system (i.e., servers and orbit) empty, ends at the first service completion epoch at which the system becomes empty again. The distribution of the maximum queue length during a busy period in the $M/M/c$ retrial queue can be numerically evaluated from Algorithm 1 of [4]. More concretely, Artalejo et al. [4] reduce the problem to computation of certain absorption probabilities that constitute the unique solution of a block-tridiagonal linear system obtained by employing first principles based on first-step analysis; as a related work, see [3] where the focus is on level-dependent QBD processes. The maximum queue length appears to be a performance descriptor of practical relevance in retrial queues since it allows us to deal with queueing systems that do not necessarily operate in stationary regime; see e.g. [3, Sections 3 and 4]. Artalejo et al. [3] apply the *busy-period* version $Z(T)$ of the maximum queue length to two problems arising from call center management. To be concrete, they consider the retrial queueing system proposed by Aguir et al. [1] to model a call center operating under the simultaneously presence of customer balking and retrials due to impatience, as well as the external-rule system investigated by Masi et al. [19] to formulate a routing rule for a resource-sharing call center.

Our major motivation is to complement the treatment of $Z(T)$ in [3,4] by presenting the random variable $Z(t_0)$ as an important descriptor that does not assume steady-state conditions. In order to specify conditions under which $Z(t_0)$ is preferred to its busy-period counterpart $Z(T)$, we point out that these conditions mainly concern with the selection of the epoch t_0 and the traffic load ρ. More concretely, the time interval $[0, t_0]$ under consideration is suggested to be comparatively smaller, on average, than the random-length interval $[0, T]$ describing a busy period; otherwise, it is reasonable that the asymptotic distribution of $Z(t_0)$ converges weakly to that of $Z(T)$ as t_0 tends to the expected length of a busy period, which is finite if and only if $\rho < 1$. In this sense, it is worth noting that the busy-period version may yield a defective distribution as the bivariate process $\mathcal{X}$ is not positive recurrent; i.e., $\rho \geqslant 1$. Not surprisingly, the distribution of the maximum queue length during a fixed time interval is always non-defective, which may play a key role for practical use. For example, a practically relevant problem in call center management is the optimal selection of the number c of agents. The launching of a new product (or the sale of a season product) commonly implies that, for a better quality of service, special resources are assigned during a period of time with fixed length t_0. The role of $Z(T)$ in [3,4] is then suggested to be replaced by the random variable $Z(t_0)$ since the underlying analysis requires transient dynamics, rather than the steady state.

The paper is organized as follows. In Section 2, we characterize the distribution of the maximum queue length during a certain time interval $[0, t_0]$ in matrix exponential form. In Section 3, we use splitting methods to compute the solution. We analyze an algorithmic solution defined in terms of the way the exponential of a matrix is appropriately computed from the Trotter product formula; see [30]. Our solution is derived, as accurately as possible, by using bounds based on the norm induced by the l_∞ vector norm. The accuracy of our solution is discussed in Section 4. Finally, concluding remarks are given in Section 5.

2. A solution in matrix exponential form

Let $Z(t_0)$ be the maximum queue length during the interval $[0, t_0]$, for a fixed time epoch $t_0 > 0$, which amounts to $\sup\{C(t) + N(t) : 0 \leqslant t \leqslant t_0\}$. We assume from now on that $X(0) = (i,j) \in \mathcal{S}$, and evaluate $P(Z(t_0) \leqslant x | X(0) = (i,j))$ for values $x \geqslant i + j$, since $P(Z(t_0) \leqslant x | X(0) = (i,j)) = 0$ if $x < i + j$.

For each value $x \geqslant i + j$, we consider an absorbing CTMC $\overline{\mathcal{X}}(x) = \{\overline{X}(t) : t \geqslant 0\}$ defined on the state space

$$\overline{\mathcal{S}}(x) = \bigcup_{j'=0}^{x} l(j') \cup \{x+1\},$$

where the subset $\cup_{j'=0}^{x} l(j')$ is an irreducible class of transient states, and the state $x+1$ is obtained by lumping the states of $\cup_{j'=x+1}^{\infty} l(j')$ to make a single absorbing state. The infinitesimal generator of $\overline{\mathcal{X}}(x)$ has the form

$$\overline{\mathbf{Q}}(x) = \begin{pmatrix} \mathbf{T}(x) & \mathbf{t}_{x+1}(x) \\ \mathbf{0}_{J(x)}^T & 0 \end{pmatrix}, \tag{4}$$

where $J(x)$ is the cardinality of the class $\cup_{j'=0}^{x} l(j')$ of transient states, $\mathbf{0}_k$ is the column vector of order k of 0s, and ^{T} is the symbol of transposition. The column vector $\mathbf{t}_{x+1}(x)$ and the matrix $\mathbf{T}(x)$ are given by

$$\mathbf{t}_{x+1}(x) = \begin{pmatrix} \mathbf{0}_{J(x-1)} \\ \lambda \mathbf{e}_{1+\min\{x,c\}} \end{pmatrix}, \quad \mathbf{T}(x) = \begin{pmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,1} & & & \\ \mathbf{A}_{1,0} & \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{A}_{x-1,x-2} & \mathbf{A}_{x-1,x-1} & \mathbf{A}_{x-1,x} \\ & & & \mathbf{A}_{x,x-1} & \mathbf{A}_{x,x} \end{pmatrix},$$

where $\mathbf{e}_k$ is the column vector of order k of 1s.

We may observe that, for $x \geqslant i + j$, the conditional probability $P(Z(t_0) > x | X(0) = (i,j))$ is equivalent to the probability that, starting from the initial state $\overline{X}(0) = (i,j)$, the process $\overline{\mathcal{X}}(x)$ visits the absorbing state $x+1$ at time t_0. Based on the structured form of $\overline{\mathbf{Q}}(x)$ in (4), we have that the standard transition function $\overline{\mathbf{P}}(t_0; x)$ (with entries $P(\overline{X}(t_0) = (i',j') | \overline{X}(0) = (i,j))$, for $(i,j), (i',j') \in \overline{\mathcal{S}}(x)$) of the process $\overline{\mathcal{X}}(x)$ can be expressed as

$$\overline{\mathbf{P}}(t_0; x) = \begin{pmatrix} \overline{\mathbf{P}}^*(t_0; x) & \overline{\mathbf{p}}_{x+1}(t_0; x) \\ \mathbf{0}_{J(x)}^T & 1 \end{pmatrix}.$$

As a result, we write down

$$P(Z(t_0) > x | X(0) = (i,j)) = \overline{\mathbf{e}}_{J(x)}(i,j) \overline{\mathbf{p}}_{x+1}(t_0; x),$$

where $\overline{\mathbf{e}}_{J(x)}(i,j)$ is a row vector of order $J(x)$ such that all its entries equal 0, except for the entry associated with the state (i,j) which equals 1. Since $\overline{\mathbf{P}}(t_0; x) = \exp\{\overline{\mathbf{Q}}(x) t_0\}$, straightforward algebra leads us to

$$\overline{\mathbf{p}}_{x+1}(t_0; x) = \left(\mathbf{I}_{J(x)} - \exp\{\mathbf{T}(x) t_0\}\right)(-\mathbf{T}^{-1}(x)) \mathbf{t}_{x+1}(x), \tag{5}$$

where $\mathbf{I}_k$ denotes the identity matrix of order k. We may remark that the matrix $-\mathbf{T}^{-1}(x)$ in (5) is non-negative for every value $x \geqslant 0$, since it consists of the expected total times spent in each transient state (i',j') during the time until absorption, starting from the initial state (i'',j''), with $(i',j'),(i'',j'') \in \cup_{x'=0}^{x} l(x')$. Indeed, the real part of each eigenvalue of $\mathbf{T}(x)$ is strictly negative, the matrix $\mathbf{T}(x)$ is stable and, consequently, $\exp\{\mathbf{T}(x)u\}$ tends to $\mathbf{0}_{J(x)\times J(x)}$ as $u \to \infty$, where $\mathbf{0}_{k\times k'}$ is the null matrix of dimension $k \times k'$. The exponential of the matrix $\mathbf{T}(x)u$ records the probabilities that up to time u the process $\overline{\mathcal{X}}(x)$ does not leave the class $\cup_{j'=0}^{x} l(j')$ of transient states, so that $\int_0^\infty \exp\{\mathbf{T}(x)u\}\mathbf{t}_{x+1}(x)du = (-\mathbf{T}^{-1}(x))\mathbf{t}_{x+1}(x)$ contains the conditional probabilities that the absorption into $x+1$ occurs in a finite time. Since $\cup_{j'=0}^{x} l(j')$ is a finite class of transient states, we have that $(-\mathbf{T}^{-1}(x))\mathbf{t}_{x+1}(x) = \mathbf{e}_{J(x)}$.

This has the following immediate consequence.

Theorem 1. *For initial numbers $i \in \{0,1,\ldots,c\}$ of busy servers and $j \in \mathbb{N}_0$ of customers in orbit, the distribution of the maximum queue length during the interval $[0,t_0]$ is specified by $P(Z(t_0) \leqslant x | X(0) = (i,j)) = 0$ if $x < i+j$, and $1 - \overline{\mathbf{e}}_{J(x)}(i,j)(\mathbf{I}_{J(x)} - \exp\{\mathbf{T}(x)t_0\})\mathbf{e}_{J(x)}$ if $x \geqslant i+j$.*

In principle, the matrix exponential $\exp\{\mathbf{T}(x)t_0\}$ in Theorem 1 could be computed in many ways; for a review of the main procedures, see [6,20]. However, consideration of computational stability and efficiency indicates that some of these procedures are preferable to others, but that none are completely satisfactory when they are implemented as general-purpose algorithms.

3. Computation of the matrix exponential solution

In this section, we present an approach based on splitting methods and eigenvalues/eigenvectors to compute the matrix exponential $\exp\{\mathbf{T}(x)t_0\}$ in an iterative manner. The main idea in our approach is closely related to the Trotter product formula:

$$\exp\{\mathbf{T}(x)t_0\} = \lim_{p\to\infty}(\exp\{\mathbf{U}(x)t\}\exp\{\mathbf{V}(x)t\})^p, \tag{6}$$

where $t = p^{-1}t_0$, for a certain splitting $\mathbf{T}(x) = \mathbf{U}(x) + \mathbf{V}(x)$; see [30]. Note that (6) becomes $\exp\{\mathbf{T}(x)t_0\} = \exp\{\mathbf{U}(x)t_0\}\exp\{\mathbf{V}(x)t_0\}$ if $\mathbf{U}(x)$ and $\mathbf{V}(x)$ commute. Moller and van Loan [20] suggest to use the approximation

$$\exp\{\mathbf{T}(x)t_0\} \simeq (\exp\{\mathbf{U}(x)t\}\exp\{\mathbf{V}(x)t\})^{p_0}, \tag{7}$$

for an appropriately selected integer p_0 and $t = p_0^{-1}t_0$. This approximation is of particular interest when the exponential of the matrices $\mathbf{U}(x)t$ and $\mathbf{V}(x)t$ can be explicitly and efficiently computed, which is our case here. This allows us to indicate that, under a simple condition on the service and retrial rates, the approach derived from (13) and (14) is preferable to other methods in the literature [20].

For the $M/M/c$ retrial queue, we suggest to use the following choices for $\mathbf{U}(x)$ and $\mathbf{V}(x)$:

(a) For $1 \leqslant x \leqslant c$,

$$\mathbf{U}(x) = \begin{pmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,1} & & & \\ & \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & & \\ & & \ddots & \ddots & \\ & & & \mathbf{A}_{x-1,x-1} & \mathbf{A}_{x-1,x} \\ & & & & \mathbf{A}_{x,x} \end{pmatrix}, \quad \mathbf{V}(x) = \begin{pmatrix} 0 & & & & \\ \mathbf{A}_{1,0} & \mathbf{0}_{2\times 2} & & & \\ & \mathbf{A}_{2,1} & \mathbf{0}_{3\times 3} & & \\ & & \ddots & \ddots & \\ & & & \mathbf{A}_{x,x-1} & \mathbf{0}_{(x+1)\times(x+1)} \end{pmatrix}.$$

(b) For $c+1 \leqslant x$,

$$\mathbf{U}(x) = \begin{pmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,1} & & & & & \\ & \ddots & \ddots & & & & \\ & & \mathbf{A}_{c,c} & \mathbf{A}_{c,c+1} & & & \\ & & & \mathbf{A}^*_{c+1,c+1} & \mathbf{A}_{c+1,c+2} & & \\ & & & & \ddots & \ddots & \\ & & & & & \mathbf{A}^*_{x-1,x-1} & \mathbf{A}_{x-1,x} \\ & & & & & & \mathbf{A}^*_{x,x} \end{pmatrix}, \quad \mathbf{V}(x) = \begin{pmatrix} 0 & & & & & & \\ \mathbf{A}_{1,0} & \mathbf{0}_{2\times 2} & & & & & \\ & \ddots & \ddots & & & & \\ & & \mathbf{A}_{c,c-1} & \mathbf{0}_{(c+1)\times(c+1)} & & & \\ & & & \mathbf{A}_{c+1,c} & \hat{\mathbf{A}}^*_{c+1,c+1} & & \\ & & & & \ddots & \ddots & \\ & & & & & \mathbf{A}_{x,x-1} & \hat{\mathbf{A}}^*_{x,x} \end{pmatrix},$$

where $\hat{\mathbf{A}}^*_{y,y}$ and $\mathbf{A}^*_{y,y}$ are square matrices of order $c+1$ with elements $(\hat{\mathbf{A}}^*_{y,y})_{k,k'} = (y-c)\mu$ if $k = c+1$ and $k' = c+1$, and 0 otherwise, and $(\mathbf{A}^*_{y,y})_{k,k'} = (y-k+1)\mu$ if $1 \leqslant k \leqslant c$ and $k' = k+1$, $-(\lambda + (k-1)\nu + (y-k+1)\mu)$ if $1 \leqslant k \leqslant c+1$ and $k' = k$, and 0 otherwise.

Our discussion in Theorem 2 is strongly based on conditions for diagonalization of the matrix $\mathbf{U}(x)$ with $x \geqslant 1$ in terms of eigenvalues and eigenvectors.

Lemma 1.

(i) *For each value $x \geqslant 1$, a sufficient condition for the matrix $\mathbf{U}(x)$ to be diagonalized is $\nu \neq \mu$ and*

$$\nu \neq \left(1 + \frac{y' - y}{l - l'}\right)\mu, \tag{8}$$

for every pair (y, y') of integers with $0 \leqslant y < y' \leqslant x$ and integers $l \neq l'$ with $0 \leqslant l \leqslant \min\{y, c\}$ and $0 \leqslant l' \leqslant \min\{y', c\}$.

(ii) *Under the assumption that $\nu \neq \mu$, the $1 + y$ (respectively, $1 + c$) eigenvalues $r(y; l) = -(\lambda + l\nu + (y - l)\mu)$, for $0 \leqslant l \leqslant \min\{y, c\}$, of the sub-matrices $\mathbf{A}_{y,y}$ (respectively, $\mathbf{A}^*_{y,y}$) are distinct. Moreover, we can specify left and right eigenvectors, respectively denoted by $\mathbf{w}(y; l) = (w_{l'}(y; l))$ and $\mathbf{v}(y; l) = (v_{l'}(y; l))$, associated with the eigenvalue $r(y; l)$ from the entries*

$$w_{l'}(y; l) = \begin{cases} 0, & \text{if } 0 \leqslant l' \leqslant l - 1, \\ 1, & \text{if } l' = l, \\ \binom{y-l}{y-l'}\left(\frac{\nu}{\mu} - 1\right)^{-(l'-l)}, & \text{if } l + 1 \leqslant l' \leqslant \min\{y, c\}, \end{cases} \tag{9}$$

$$v_{l'}(y; l) = \begin{cases} \binom{y-l'}{y-l}\left(1 - \frac{\nu}{\mu}\right)^{-(l-l')}, & \text{if } 0 \leqslant l' \leqslant l - 1, \\ 1, & \text{if } l' = l, \\ 0, & \text{if } l + 1 \leqslant l' \leqslant \min\{y, c\}. \end{cases} \tag{10}$$

Proof. Since the sub-matrices $\mathbf{A}_{y,y}$ with $0 \leqslant y \leqslant c$, and $\mathbf{A}^*_{y,y}$ with $c + 1 \leqslant y \leqslant x$, have block bi-diagonal form, their eigenvalues are given by their diagonal elements $r(y; l) = -(\lambda + l\nu + (y - l)\mu)$, for $0 \leqslant l \leqslant \min\{y, c\}$. Then, for every fixed $y \geqslant 0$, the eigenvalues $r(y; 0), r(y; 1), \ldots, r(y; y)$ (respectively, $r(y; 0), r(y; 1), \ldots, r(y; c)$) of the sub-matrix $\mathbf{A}_{y,y}$ (respectively, $\mathbf{A}^*_{y,y}$) are distinct if and only if $\nu \neq \mu$. On the other hand, for every pair (y, y') of integers with $0 \leqslant y < y' \leqslant x$, the eigenvalue $r(y; l)$ of a first sub-matrix $\mathbf{A}_{y,y}$ or $\mathbf{A}^*_{y,y}$, and the eigenvalue $r(y'; l')$ of a second sub-matrix $\mathbf{A}_{y',y'}$ or $\mathbf{A}^*_{y',y'}$ are distinct if (8) holds, from which it follows that eigenvalues of two different sub-matrices of the family $\{\mathbf{A}_{0,0}, \mathbf{A}_{1,1}, \ldots, \mathbf{A}_{x,x}\}$ if $1 \leqslant x \leqslant c$, or the family $\{\mathbf{A}_{0,0}, \ldots, \mathbf{A}_{c,c}, \mathbf{A}^*_{c+1,c+1}, \ldots, \mathbf{A}^*_{x,x}\}$ if $c + 1 \leqslant x$, are distinct if and only if the rates ν and μ satisfy (8) for every pair (y, y') of integers with $0 \leqslant y < y' \leqslant x$ and integers $l \neq l'$ with $0 \leqslant l \leqslant \min\{y, c\}$ and $0 \leqslant l' \leqslant \min\{y', c\}$.

To derive (9) for $1 \leqslant y \leqslant c$, we express the matrix equation $\mathbf{w}(y; l)\mathbf{A}_{y,y} = r(y; l)\mathbf{w}(y; l)$ as

$$(r(y; l) - r(y; 0))w_0(y; l) = 0,$$
$$(r(y; l) - r(y; l'))w_{l'}(y; l) = (y - l' + 1)\mu w_{l'-1}(y; l), \quad 1 \leqslant l' \leqslant y.$$

If $\nu \neq \mu$, then $r(y; l) \neq r(y; l')$ for every integer $l' \in \{0, 1, \ldots, l - 1\}$ and, consequently, $w_0(y; l) = w_1(y; l) = \cdots = w_{l-1}(y; l) = 0$. Then, the above equations result in

$$w_{l'}(y; l) = \prod_{k=l}^{l'-1} \frac{(y - k)\mu}{r(y; l) - r(y; k + 1)} w_l(y; l), \quad l + 1 \leqslant l' \leqslant y,$$

which completes the proof of (9) if we select the value $w_l(y; l) = 1$. Eq. (10) in the case $1 \leqslant y \leqslant c$ is readily obtained from the matrix equation $\mathbf{A}_{y,y}\mathbf{v}(y; l) = r(y; l)\mathbf{v}(y; l)$. Expressions for the entries of $\mathbf{w}(y; l)$ in (9) and $\mathbf{v}(y; l)$ in (10), for $c + 1 \leqslant x$, are similarly derived by replacing $\mathbf{A}_{y,y}$ by $\mathbf{A}^*_{y,y}$ in the preceding arguments. □

In what follows, we show that the matrix exponential $\exp\{\mathbf{T}(x)t_0\}$ in (5) can be evaluated in an efficient manner by means of the exponential of the matrices $\mathbf{U}(x)t$ and $\mathbf{V}(x)t$. To be concrete, Theorem 2 establishes that $\exp\{\mathbf{U}(x)t\}$ can be iteratively computed starting with the term $\exp\{\mathbf{U}(0)t\} = e^{-\lambda t}$, under the assumption of Lemma 1. (i). Unlike the iterative scheme for $\exp\{\mathbf{U}(x)t\}$ in Theorem 2, our expression for $\exp\{\mathbf{V}(x)t\}$ in Theorem 3 does not involve any condition associated with the eigenvalues and/or eigenvectors of the sub-matrices $\mathbf{A}_{y,y}$ and $\mathbf{A}^*_{y,y}$, Hence, Eq. (14) in Theorem 3 is valid in full generality, that is, even if the rates ν and μ do not satisfy Lemma 1. (i).

To begin with, we notice that $\mathbf{U}(x)$ is a block bi-diagonal matrix, whence its eigenvalues are given by the eigenvalues of the sub-matrices $\{\mathbf{A}_{0,0}, \mathbf{A}_{1,1}, \ldots, \mathbf{A}_{x,x}\}$ if $1 \leqslant x \leqslant c$, and $\{\mathbf{A}_{0,0}, \ldots, \mathbf{A}_{c,c}, \mathbf{A}^*_{c+1,c+1}, \ldots, \mathbf{A}^*_{x,x}\}$ if $c + 1 \leqslant x$. As a result, for a fixed value $x \geqslant i + j$, Lemma 1. (i) implies that the eigenvalues $\{r(y; l) = -(\lambda + l\nu + (y - l)\mu) : 0 \leqslant l \leqslant \min\{y, c\}, 0 \leqslant y \leqslant x\}$ of the matrix $\mathbf{U}(x)$ are distinct, and $\mathbf{U}(x)$ can be diagonalized since it possesses $J(x)$ linearly independent eigenvectors. Specifically, we may write that $\mathbf{U}(x) = \mathbf{R}(x)\mathbf{E}(x)\mathbf{R}^{-1}(x)$, where $\mathbf{E}(x)$ is a diagonal matrix whose diagonal elements are given by the eigenvalues $\{r(y; l) = -(\lambda + l\nu + (y - l)\mu) : 0 \leqslant l \leqslant \min\{y, c\}, 0 \leqslant y \leqslant x\}$ of $\mathbf{U}(x)$, and the columns of $\mathbf{R}(x)$ correspond to the right eigenvectors of $\mathbf{U}(x)$ associated with these eigenvalues. From Lemma 1. (ii), it is seen that $\mathbf{R}(x)$ can be expressed as

$$\mathbf{R}(x) = \begin{pmatrix} \mathbf{R}_{0,0} & \mathbf{R}_{0,1} & \mathbf{R}_{0,2} & \cdots & \mathbf{R}_{0,x-1} & \mathbf{R}_{0,x} \\ & \mathbf{R}_{1,1} & \mathbf{R}_{1,2} & \cdots & \mathbf{R}_{1,x-1} & \mathbf{R}_{1,x} \\ & & \mathbf{R}_{2,2} & \cdots & \mathbf{R}_{2,x-1} & \mathbf{R}_{2,x} \\ & & & \ddots & & \vdots \\ & & & & \mathbf{R}_{x-1,x-1} & \mathbf{R}_{x-1,x} \\ & & & & & \mathbf{R}_{x,x} \end{pmatrix}, \quad (11)$$

where $\mathbf{R}_{y,y}$ is a square matrix of order $1+\min\{y,c\}$ and its columns are the right eigenvectors $\mathbf{v}(y;l)$, for $0 \leqslant l \leqslant \min\{y,c\}$, in (10). The columns of the sub-matrix $\mathbf{R}_{y',y} = (\mathbf{r}(y',y;0), \mathbf{r}(y',y;1), \ldots, \mathbf{r}(y',y;\min\{y,c\}))$ are specified as

$$\mathbf{r}(y',y;l) = \prod_{k=y'}^{y-1} \left(r(y;l)\mathbf{I}_{1+\min\{k,c\}} - \mathbf{B}_{k,k}\right)^{-1} \mathbf{A}_{k,k+1}\mathbf{v}(y;l),$$

with $\mathbf{B}_{k,k} = \mathbf{A}_{k,k}$ if $0 \leqslant k \leqslant c$, and $\mathbf{A}^*_{k,k}$ if $c+1 \leqslant k$. Note that, by [17, page 150], the matrices $\left(r(y;l)\mathbf{I}_{1+\min\{k,c\}} - \mathbf{B}_{k,k}\right)^{-1}$ are given by

$$\left(r(y;l)\mathbf{I}_{1+\min\{k,c\}} - \mathbf{B}_{k,k}\right)^{-1} = \sum_{l'=0}^{\min\{k,c\}} \frac{\mathbf{v}(k;l')\mathbf{w}(k;l')}{r(y;l) - r(k;l')},$$

where the left and right eigenvectors $\mathbf{w}(k;l')$ and $\mathbf{v}(k;l')$ are evaluated from (9) and (10), respectively.

The identity $\mathbf{U}(x) = \mathbf{R}(x)\mathbf{E}(x)\mathbf{R}^{-1}(x)$ implies

$$\exp\{\mathbf{U}(x)t\} = \mathbf{R}(x)\mathbf{E}(x;t)\mathbf{R}^{-1}(x), \quad (12)$$

where $\mathbf{E}(x;t) = \mathrm{diag}(e^{r(0;0)t}, e^{r(1;0)t}, e^{r(1;1)t}, \ldots, e^{r(x;0)t}, \ldots, e^{r(x;\min\{x,c\})t})$, and $r(y;l) = -(\lambda + l\nu + (y-l)\mu)$, for $0 \leqslant l \leqslant \min\{y,c\}$ and $0 \leqslant y \leqslant x$. If we express the matrices $\mathbf{R}(x)$ and $\mathbf{E}(x;t)$ in structured form as

$$\mathbf{R}(x) = \begin{pmatrix} \mathbf{R}(x-1) & \mathbf{N}(x) \\ \mathbf{0}_{(1+\min\{x,c\})\times J(x-1)} & \mathbf{R}_{x,x} \end{pmatrix},$$

and $\mathbf{E}(x;t) = \mathrm{diag}(\mathbf{E}(x-1;t), \mathbf{D}_x(t))$, then we can reduce the computation in (12) to previously computed sub-matrices.

Theorem 2. *Let us assume that the rates* ν *and* μ *satisfy* Lemma 1. (i). *Then, starting from* $\exp\{\mathbf{U}(0)t\} = e^{-\lambda t}$, *the matrix exponential* $\exp\{\mathbf{U}(x)t\}$ *can be computed as*

$$\exp\{\mathbf{U}(x)t\} = \begin{pmatrix} \exp\{\mathbf{U}(x-1)t\} & \mathbf{N}(x;t) \\ \mathbf{0}_{(1+\min\{x,c\})\times J(x-1)} & \exp\{\mathbf{B}_{x,x}t\} \end{pmatrix}, \quad x \geqslant 1, \quad (13)$$

with $\mathbf{N}(x;t) = (\mathbf{N}(x)\mathbf{D}_x(t) - \exp\{\mathbf{U}(x-1)t\}\mathbf{N}(x))\mathbf{R}^{-1}_{x,x}$.

It should be noted that $\exp\{\mathbf{B}_{x,x}t\} = \mathbf{R}_{x,x}\mathbf{D}_x(t)\mathbf{R}^{-1}_{x,x}$ and $\mathbf{R}^{-1}_{x,x}$ is given by

$$\mathbf{R}^{-1}_{x,x} = \begin{pmatrix} c^{-1}(x;0)\mathbf{w}(x;0) \\ c^{-1}(x;1)\mathbf{w}(x;1) \\ \vdots \\ c^{-1}(x;\min\{x,c\})\mathbf{w}(x;\min\{x,c\}) \end{pmatrix},$$

where the left eigenvectors $\mathbf{w}(x;l)$ are specified by (9) and the constant $c(x;l)$ is given by $c(x;l) = \sum_{l'=0}^{\min\{x,c\}} w_{l'}(x;l)v_{l'}(x;l)$; note that $c(x;l) = 1$ by (9) and (10).

Based on the sparse form of the matrix $\mathbf{V}(x)$, we may derive an explicit expression for its exponential. The proof of Theorem 3 is based on the fact that $\exp\{\mathbf{V}(x)t\} = \sum_{k=0}^{x}(k!)^{-1}(\mathbf{V}(x)t)^k$ if $x \leqslant c$, and an inductive argument if $c+1 \leqslant x$, and it is thus omitted. Expressions for the matrices $\mathbf{S}(y;y') = \mathbf{A}_{y,y-1}\mathbf{A}_{y-1,y-2}\ldots\mathbf{A}_{y'+1,y'}$, for $0 \leqslant y' \leqslant y-1$ and $1 \leqslant y \leqslant x$, are gathered in Appendix B.

Theorem 3. *For a fixed value x, the exponential of the matrix* $\mathbf{V}(x)t$ *has the form*

$$\exp\{\mathbf{V}(x)t\} = \begin{pmatrix} \mathbf{M}(0;0) & & & & \\ \mathbf{M}(1;0) & \mathbf{M}(1;1) & & & \\ \mathbf{M}(2;0) & \mathbf{M}(2;1) & \mathbf{M}(2;2) & & \\ \vdots & \vdots & \vdots & \ddots & \\ \mathbf{M}(x;0) & \mathbf{M}(x;1) & \mathbf{M}(x;2) & \cdots & \mathbf{M}(x;x) \end{pmatrix}, \quad (14)$$

where $\mathbf{M}(y;y) = \mathbf{I}_{1+y}$ *if* $0 \leqslant y \leqslant c$, *and* $\mathbf{M}(y;y) = diag(\mathbf{I}_c, e^{(y-c)\mu t})$ *if* $c+1 \leqslant y$. *If* $0 \leqslant y' \leqslant y-1$ *and* $1 \leqslant y \leqslant c$, *then the sub-matrices* $\mathbf{M}(y;y')$ *are given by* $\mathbf{M}(y;y') = ((y-y')!)^{-1} t^{y-y'} \mathbf{S}(y;y')$. *In the case* $0 \leqslant y' \leqslant y-1$ *and* $c+1 \leqslant y$, *the sub-matrix* $\mathbf{M}(y;y')$ *has the form*

$$\begin{pmatrix} \frac{t^{y-y'}}{(y-y')!}\mathbf{I}_c & \\ & \left(e^{(y-c)\mu t} - \sum_{k=0}^{y-y'-1} \frac{((y-c)\mu t)^k}{k!}\right)((y-c)\mu)^{y'-y} \end{pmatrix} \mathbf{S}(y;y').$$

4. The accuracy of our solution and discussion

For a fixed epoch $t_0 > 0$ and each value of $x \geqslant i+j$, expressions (13) and (14) allow us to approximate the exponential of the matrix $\mathbf{T}(x)t_0$ from (7). We next select the value of p_0 in (7) by using bounds based on the norm induced by the l_∞ vector norm.

For the splitting $\mathbf{T}(x) = \mathbf{U}(x) + \mathbf{V}(x)$ the value of p_0 in (7) can be determined from the inequality

$$||\exp\{\mathbf{T}(x)t_0\} - (\exp\{\mathbf{U}(x)t\}\exp\{\mathbf{V}(x)t\})^{p_0}||_\infty \leqslant \frac{h(x;t_0)}{2p_0}, \tag{15}$$

where the function $h(x;t_0)$ is defined in terms of

$$h(x;t_0) = ||[\mathbf{U}(x), \mathbf{V}(x)]||_\infty t_0^2 e^{(\mu_\infty(\mathbf{U}(x)) + \mu_\infty(\mathbf{V}(x))t_0}, \tag{16}$$

$||\cdot||_\infty$ and $\mu_\infty(\cdot)$ denote the *maximum row sum matrix norm* and the *logarithmic norm* of a matrix, respectively, and the commutator of two matrices is given by $[\mathbf{U}(x), \mathbf{V}(x)] = \mathbf{U}(x)\mathbf{V}(x) - \mathbf{V}(x)\mathbf{U}(x)$. The proof of (15) mostly repeats arguments of Theorem 5 of [20], and it is thus omitted. We recall that the $||\cdot||_\infty$-norm of a square matrix $\mathbf{W} = (w_{i,j})$ of order k is defined by $||\mathbf{W}||_\infty = \max_{1\leqslant i\leqslant k}\sum_{j=1}^k |w_{i,j}|$, and it can be seen as the norm induced by the l_∞ vector norm; see [16]. The logarithmic norm of the matrix $\mathbf{W}$ is specified by $\mu_\infty(\mathbf{W}) = \lim_{h\to 0+} h^{-1}(||\mathbf{I}_k - h\mathbf{W}||_\infty - 1)$, and it can be evaluated by [29] as $\mu_\infty(\mathbf{W}) = \max_{1\leqslant i\leqslant k}\left\{w_{i,i} + \sum_{j\neq i}|w_{i,j}|\right\}$.

By (15), we may choose for any $\varepsilon > 0$ the value of p_0 as the first positive integer such that $(2\varepsilon)^{-1}h(x;t_0) < p_0$, which implies that $||\exp\{\mathbf{T}(x)t_0\} - (\exp\{\mathbf{U}(x)t\}\exp\{\mathbf{V}(x)t\})^{p_0}||_\infty < \varepsilon$. This selection means that, for a predetermined $\varepsilon > 0$, the integer p_0 satisfying $(2\varepsilon)^{-1}h(x;t_0) < p_0$ depends on t_0 and $x \geqslant i+j$.

In evaluating the function $h(x;t_0)$ in (16), it is seen that $\mu_\infty(\mathbf{U}(x)) = 0$, for $x \geqslant 1$, and

$$\mu_\infty(\mathbf{V}(x)) = \begin{cases} x\nu, & \text{if } 1 \leqslant x \leqslant c, \\ c\nu + (x-c)\mu, & \text{if } c+1 \leqslant x. \end{cases}$$

For $1 \leqslant x \leqslant c$, the value of $||[\mathbf{U}(x), \mathbf{V}(x)]||_\infty$ is given by

$$||[\mathbf{U}(x), \mathbf{V}(x)]||_\infty = \begin{cases} x\nu\mu, & \text{if } \lambda+\nu \leqslant \mu, \\ x\nu(\lambda+\nu), & \text{if } \lambda+\nu > \mu. \end{cases}$$

In the case $c+1 \leqslant x$, we may express

$$||[\mathbf{U}(x), \mathbf{V}(x)]||_\infty = \max\{g_1(x), g_2(x), g_3(x), g_4(x), g_5(x)\},$$

where

$$\begin{aligned} g_1(x) &= (\lambda + (c-1)\nu + (x-c)\mu)\nu + (x-c)(\lambda + (x-c-1)\mu)\mu, \\ g_2(x) &= (2\lambda + \nu + (x-c-1)\mu)c\nu + \lambda\mu, \\ g_3(x) &= \nu f(c-1) + (x-c+1)(x-c)\mu^2, \\ g_4(x) &= \nu f(c) + (c-1)(x-c)\nu\mu, \end{aligned}$$

and $g_5(x) = \nu f(0)$ if $\lambda+\nu \leqslant \mu$, and $\nu f(c-2)$ if $\lambda+\nu > \mu$, with $f(l) = x\mu + (\lambda+\nu-\mu)l$.

It is worth mentioning that, for each value $x \geqslant i+j$, our expression for the probability $P(Z(t_0) \leqslant x | X(0) = (i,j))$ in Theorem 1 requires an accurate estimation of a single row of the matrix exponential $\exp\{\mathbf{T}(x)t_0\}$, which is related to the initial numbers of i busy servers and j customers in orbit. In our approach, we suggest to approximate $P(Z(t_0) \leqslant x | X(0) = (i,j))$ by the value $p(x;t_0) \equiv p(x;t_0;\varepsilon)$ satisfying

$$p(x;t_0) = 1 - \overline{\mathbf{e}}_{J(x)}(i,j)\left(\mathbf{I}_{J(x)} - (\exp\{\mathbf{U}(x)t\}\exp\{\mathbf{V}(x)t\})^{p_0}\right)\mathbf{e}_{J(x)},$$

with $t = p_0^{-1}t_0$. This means that, for any $\varepsilon > 0$, we have

$$|P(Z(t_0) \leqslant x | X(0) = (i,j)) - p(x;t_0)| < \varepsilon, \tag{17}$$

since p_0 satisfies $(2\varepsilon)^{-1}h(x;t_0) < p_0$, and

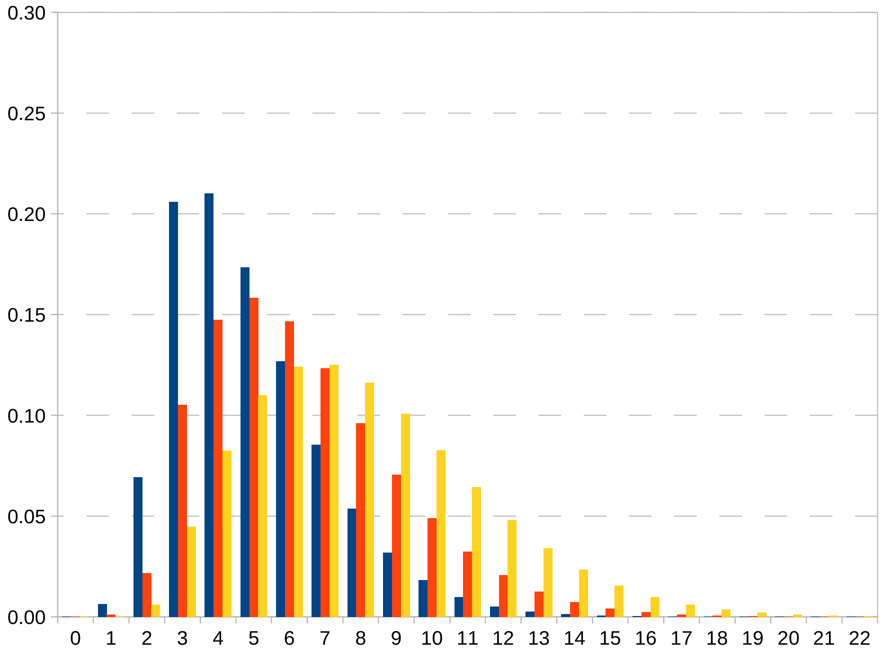

Fig. 1. The mass function of $Z(t_0)$ versus ρ for retrial queues with $c = 3$, $\nu = 3.0\sqrt{2.0}$ and $\mu = 2.5$. Values of the traffic load: $\rho = 0.8, 1.0$ and 1.2 (*from left to right*); interval length: $t_0 = 1.0$; initial state: $(i,j) = (0,0)$.

$$\begin{aligned}|P(Z(t_0) \leqslant x|X(0) = (i,j)) - p(x;t_0)| &\leqslant \bar{\mathbf{e}}_{J(x)}(i,j)|\exp\{\mathbf{T}(x)t_0\} - (\exp\{\mathbf{U}(x)t\}\exp\{\mathbf{V}(x)t\})^{p_0}|\mathbf{e}_{J(x)} \\ &\leqslant ||\exp\{\mathbf{T}(x)t_0\} - (\exp\{\mathbf{U}(x)t\}\exp\{\mathbf{V}(x)t\})^{p_0}||_\infty \\ &\leqslant \frac{h(x;t_0)}{2p_0}.\end{aligned}$$

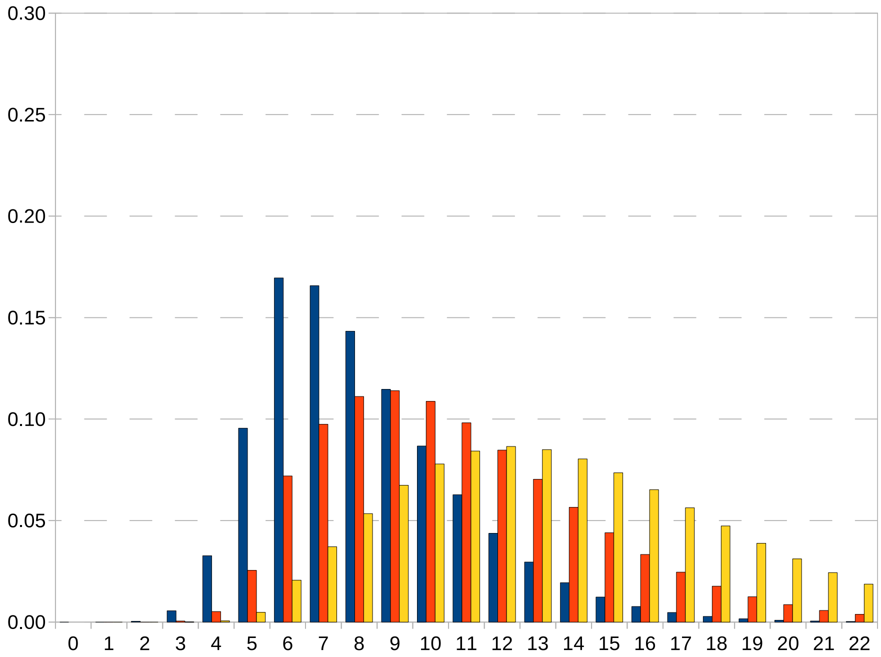

Fig. 2. The mass function of $Z(t_0)$ versus ρ for retrial queues with $c = 6$, $\nu = 3.0\sqrt{2.0}$ and $\mu = 2.5$. Values of the traffic load: $\rho = 0.8, 1.0$ and 1.2 (*from left to right*); interval length: $t_0 = 1.0$; initial state: $(i,j) = (0,0)$.

The *uniformization* method may be used, in principle, to compute the cumulative distribution function of $Z(t_0)$; see e.g. [18, Section 2.8]. More concretely, we may first uniformize the absorbing CTMC $\overline{\mathcal{X}}(x)$ with infinitesimal generator $\overline{\mathbf{Q}}(x)$ (Eq. (4)) by choosing

$$c(x) = \max\{q_{(i',j')} : (i',j') \in \overline{\mathcal{S}}(x)\},$$

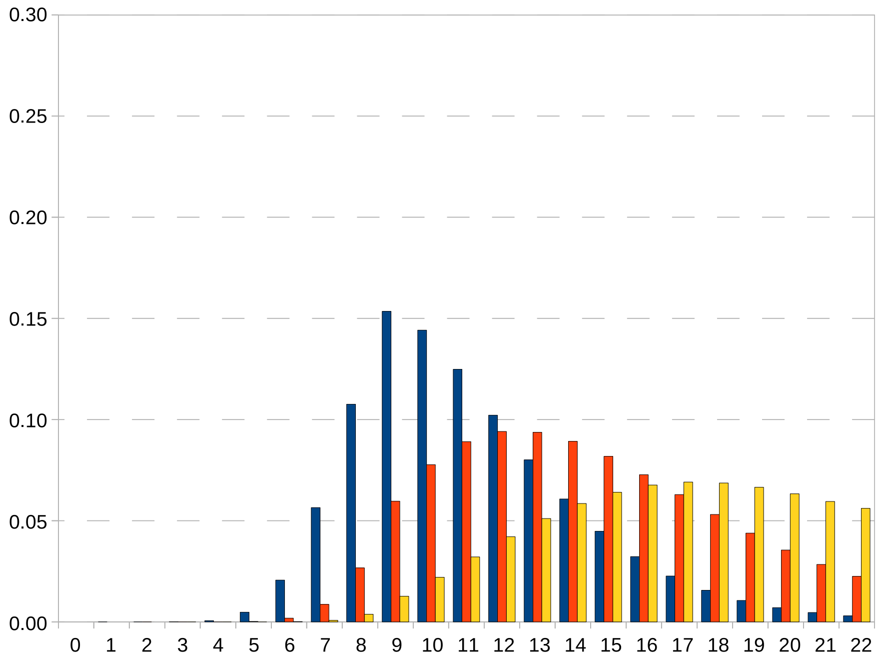

Fig. 3. The mass function of $Z(t_0)$ versus ρ for retrial queues with $c = 9$, $v = 3.0\sqrt{2.0}$ and $\mu = 2.5$. Values of the traffic load: $\rho = 0.8, 1.0$ and 1.2 (*from left to right*); interval length: $t_0 = 1.0$; initial state: $(i,j) = (0,0)$.

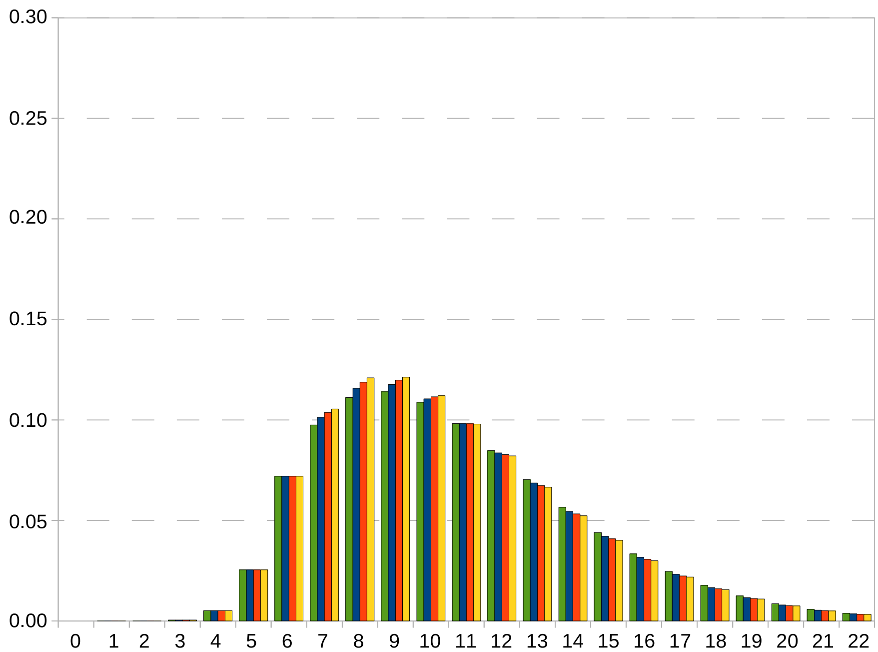

Fig. 4. The mass function of $Z(t_0)$ versus μ for retrial queues with $c = 6$, $v = 3.0\sqrt{2.0}$ and $\rho = 1.0$. Values of the retrial rate: $\mu = 2.5, 5.0, 7.5$ and 10.0 (*from left to right*); interval length: $t_0 = 1.0$; initial state: $(i,j) = (0,0)$.

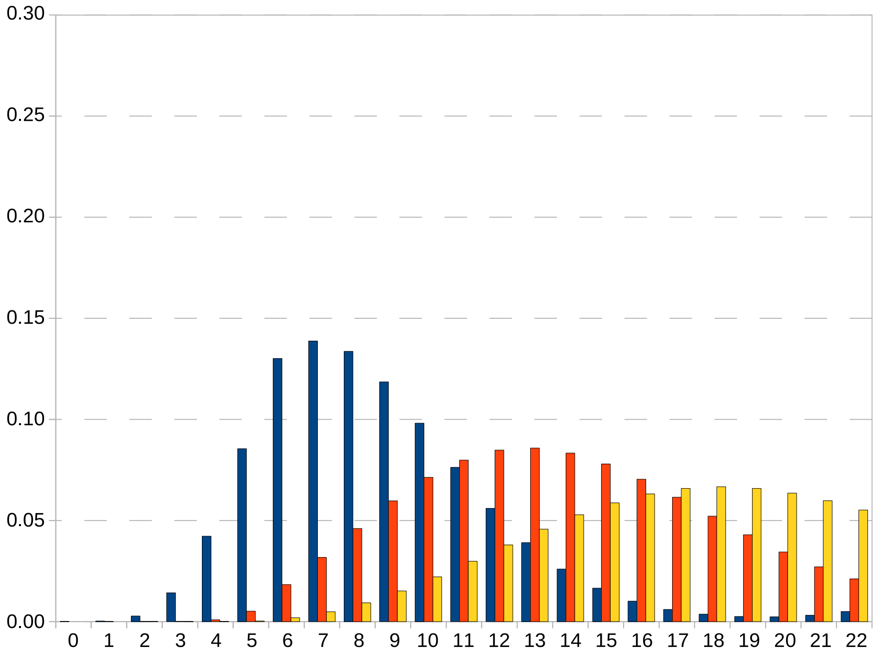

Fig. 5. The mass function of $Z(t_0)$ versus t_0 for retrial queues with $c = 6, \nu = 3.0\sqrt{2.0}$ and $\mu = 2.5$. Traffic load: $\rho = 1.5$; values of the interval length: $t_0 = 1/3, 2/3$, and 1.0 (*from left to right*); initial state: $(i,j) = (0,0)$.

and then derive a simple algorithm to compute $P(Z(t_0) \leqslant x | X(0) = (i,j))$ for values $x \geqslant i + j$. Unfortunately, it is not a very satisfactory result for our problem since there is no clear relationship between the underlying matrices associated with an application of [18, Fig. 2.4] to successive values x and $x + 1$. This drawback is closely related to the fact that $c(x)$ increases with increasing values of x, and it converges to infinity as x tends to infinity. Even for values $x \geqslant i + j$, the uniformization method might require some care if $t_0 c(x)$ should happen to be large.

For illustrative purposes, we present a few numerical results in Figs. 1–5, where we use $\varepsilon = 10^{-3}$ in (17), and we fix the initial state $(i,j) = (0,0)$. In Figs. 1–3, we plot the mass function of $Z(t_0)$ with $t_0 = 1.0$, for retrial queues with $c \in \{3, 6, 9\}$, service rate $\nu = 3.0\sqrt{2.0}$, and retrial rate $\mu = 2.5$. In each figure, we display three histograms that correspond to the values $\rho = 0.8, 1.0$ and 1.2 (from left to right) of the traffic load. It is observed that the mass function of $Z(t_0)$ behaves as a bell-shaped function, regardless of the number c of servers and the traffic load ρ. As intuition tells us, the distribution of $Z(t_0)$ has heavier tails when ρ increases, for each fixed value of c.

In Fig. 4, we consider $c = 6, \nu = 3.0\sqrt{2.0}$ and the traffic load $\rho = 1.0$, and we include four histograms, which are related to the values $\mu = 2.5, 5.0, 7.5$ and 10.0 (from left to right) of the retrial rate. For each fixed value $x \in \{0, 1, \ldots, c\}$, the probability $P(Z(t_0) = x | X(0) = (i,j))$ is shown to be constant as a function of the retrial rate μ; note that the event $Z(t_0) \in \{0, 1, \ldots, c\}$ means that the number $N(t)$ of customers in orbit is equal to zero at every time instant $t \in [0, t_0]$. It can be also inferred that the retrial rate μ does not seem to be influential at the tail of the distribution. Fig. 5 shows the effect of the length t_0 on the mass function of $Z(t_0)$ for retrial queues with $c = 6, \nu = 3.0\sqrt{2.0}, \mu = 2.5$ and the value $\rho = 1.5$ of the traffic load. Specifically, we select interval lengths $t_0 = 0.333, 0.666$ and 1.0 (from left to right). In agreement with our expectations, the distribution of $Z(t_0)$ becomes more sparse with increasing values of the interval length t_0.

5. Conclusions

In an attempt to measure the effects of extreme values on retrial queues during a predetermined time interval, we have presented a new probabilistic descriptor, namely the maximum queue length $Z(t_0)$ during the time interval $[0, t_0]$ in the main multiserver retrial queue. For a practical use, a fixed interval $[0, t_0]$ might be related to the launching of a new product, which means that, under the distributional assumptions of the $M/M/c$ retrial model, the random variable $Z(t_0)$ allows us somehow to estimate in advance those maximum resources that, under concrete cost specifications, the community of potential customers needs to have the certain knowledge that their demand will be successfully satisfied by the corresponding call center during a predetermined time interval of length t_0.

In Theorem 1, we have derived a matrix exponential form for the probability distribution function of $Z(t_0)$ for initial numbers $i \in \{0, 1, \ldots, c\}$ of busy servers and $j \in \mathbb{N}_0$ of customers in orbit. In terms of Lemma 1. (i), we have assumed simple conditions on the rates ν and μ for the resulting matrix exponential to be amenable to numerical calculation. It is clear that our assumption in Lemma 1. (i) is a technical requirement to guarantee that the eigenvalues in the set

$\{r(y;l) = -(\lambda + l\nu + (y-l)\mu) : 0 \leqslant l \leqslant \min\{y,c\}, 0 \leqslant y \leqslant x\}$ are distinct for every $x \geqslant i+j$ (consequently, the matrix $\mathbf{U}(x)$ can be diagonalized), but it may be thought of as sufficiently general to be applied in practice, even in the case $\rho \geqslant 1$. We have suggested to approximate the matrix exponential solution of Theorem 1, as accurately as possible, by using the Trotter product formula (6). Specifically, we have first analyzed a concrete splitting proposal for the underlying matrix $\mathbf{T}(x) = \mathbf{U}(x) + \mathbf{V}(x)$, and we have then studied its practical qualities when this splitting is combined with the matrix norm $||\cdot||_\infty$ induced by the l_∞ vector norm. In our solution, Lemma 1. (i) is combined with Theorem 2, thus implying that the matrix $\exp\{\mathbf{U}(x)t\}$ can be iteratively computed, starting with $\exp\{\mathbf{U}(0)t\} = e^{-\lambda t}$, in terms of the previously-evaluated matrix $\exp\{\mathbf{U}(x-1)t\}$ and the exponential of the matrix $\mathbf{B}_{x,x}t$. More concretely, the matrix exponential $\exp\{\mathbf{B}_{x,x}t\}$ in (13) has dimension $1+\min\{x,c\}$, and admits an explicit expression in terms of the eigenvalues and left and right eigenvectors of Lemma 1. (ii). The rest of the operations in our splitting solution only involve the product of known matrices. A particularly feature of our solution is that the approach allows us to obtain global error control since the approximating solution $p(x;t_0) \equiv p(x;t_0;\varepsilon)$ verifies Eq. (17) uniformly in $x \geqslant i+j$, for any $\varepsilon > 0$ and initial numbers i of busy servers and j of customers in orbit.

It should be noted that there are two reasons to formulate the level-dependent QBD process in (3) in terms of the number of customers in the system, instead of the orbit size; see, for example, [5, Section 2.1]. First, the level $l(j) = \{(i,j-i) : 0 \leqslant i \leqslant \min\{j,c\}\}$, with $j \geqslant 0$, in Section 1 allows us to integrate both numbers $C(t)$ and $N(t)$ (i.e., busy servers and orbit size) into the maximum queue length $Z(t_0)$ in a natural manner. In contrast, the level $l'(j) = \{(i,j) : 0 \leqslant i \leqslant c\}$, with $j \geqslant 0$, in [5, Section 2.1] is related to the number $N(t)$, so additional arrangements are necessary to incorporate the number $C(t)$ of busy servers into the modeling aspects. Second, the use of the level $l(j)$ leads us to bidiagonal matrices $\mathbf{A}_{j,j}$ in (3). This means that eigenvalues and eigenvectors of $\mathbf{A}_{j,j}$ and $\mathbf{A}^*_{j,j}$ can be determined in an explicit manner. Therefore, the bidiagonal structure of $\mathbf{A}_{j,j}$ in (3) is the key step to derive the iterative procedure in (13) for computing the matrix exponential $\exp\{\mathbf{U}(x)t\}$. Note that, in the case of $l'(j)$, the tridiagonal structure of $\mathbf{A}_{j,j}$ seems to require the use of numerical methods for evaluating their eigenvalues and eigenvectors.

The matrix exponential expression in Theorem 1 is inherently connected to the well-known solution to the forward and backward equations of the absorbing CTMC $\overline{\mathcal{X}}(x)$ in (4), for $x \geqslant i+j$. Therefore, our methodology could be applied with appropriate modifications to other retrial queues characterized by an irreducible QBD process $\mathcal{X}$, such as the retrial models of Aguir et al. [1], and Masi et al. [19]; see [3, Sections 3 and 4] for an analysis based on the busy-period version $Z(T)$. To conclude, we point out that the matrix exponential form and its evaluation in terms of splitting methods can be also used when $\mathcal{X}$ is an absorbing QBD process, which is the case of the process $\mathcal{X}$ in [14]. More concretely, the process $\mathcal{X}$ in [14] records the numbers of parasites and hosts alive at an arbitrary time in an ecosystem. The authors in [14] analyze two algorithmic solutions defined in terms of the way the exponential of the resulting matrix $\mathbf{T}(x)$ is appropriately computed from the Trotter product formula, and they derive two bounds based on the spectral norm $||\cdot||_S$ and the norm $||\cdot||_\infty$. They also discuss on the use of the maximum number $Z(t_0)$ of individuals alive in an attempt to describe how a community of parasites and hosts is affected by extreme values.

Acknowledgments

The authors thank two anonymous referees for their constructive criticism of the presentation of the paper.

Appendix A. Expressions for the sub-matrices $\mathbf{A}_{j,j'}$ with $j' \in \{j-1, j, j+1\}$

In Section 1, the infinitesimal generator $\mathbf{Q}$ of the CTMC $\mathcal{X}$ consists of sub-matrices $\mathbf{A}_{j,j'}$ with $j' \in \{j-1,j,j+1\}$. For states of the jth level, a natural ordering is defined as $(0,j) \prec (1,j-1) \prec \cdots \prec (\min\{j,c\}-1, j-\min\{j,c\}+1) \prec (\min\{j,c\}, j-\min\{j,c\})$, where $a \prec b$ denotes that state a precedes state b. Then, it can be routinely verified that

$$\mathbf{A}_{j,j-1} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ \nu & 0 & \cdots & 0 & 0 \\ 0 & 2\nu & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & (j-1)\nu & 0 \\ 0 & 0 & \cdots & 0 & j\nu \end{pmatrix}, \quad 1 \leqslant j \leqslant c, \quad \mathbf{A}_{j,j-1} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 & 0 \\ \nu & 0 & \cdots & 0 & 0 & 0 \\ 0 & 2\nu & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & (c-1)\nu & 0 & 0 \\ 0 & 0 & \cdots & 0 & c\nu & 0 \end{pmatrix}, \quad c+1 \leqslant j,$$

$$\mathbf{A}_{j,j} = \begin{pmatrix} -q_{(0,j)} & j\mu & 0 & \cdots & 0 & 0 \\ 0 & -q_{(1,j-1)} & (j-1)\mu & \cdots & 0 & 0 \\ 0 & 0 & -q_{(2,j-2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -q_{(j-1,1)} & \mu \\ 0 & 0 & 0 & \cdots & 0 & -q_{(j,0)} \end{pmatrix}, \quad 0 \leqslant j \leqslant c,$$

$$\mathbf{A}_{j,j} = \begin{pmatrix} -q_{(0,j)} & j\mu & 0 & \cdots & 0 & 0 \\ 0 & -q_{(1,j-1)} & (j-1)\mu & \cdots & 0 & 0 \\ 0 & 0 & -q_{(2,j-2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -q_{(c-1,j-c+1)} & (j-c+1)\mu \\ 0 & 0 & 0 & \cdots & 0 & -q_{(c,j-c)} \end{pmatrix}, \quad c+1 \leqslant j,$$

$$\mathbf{A}_{j,j+1} = \begin{pmatrix} 0 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 0 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}, \quad 0 \leqslant j \leqslant c-1, \quad \mathbf{A}_{j,j+1} = \begin{pmatrix} 0 & \lambda & 0 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}, \quad c+1 \leqslant j.$$

Appendix B. Expressions for the sub-matrices $\mathbf{S}(y; y')$ in Theorem 3

Sub-matrices $\mathbf{S}(y; y')$ in Theorem 3 are defined by $\mathbf{A}_{y,y-1}\mathbf{A}_{y-1,y-2}\ldots\mathbf{A}_{y'+1,y'}$. They can be evaluated as follows:

(a) For $0 \leqslant y' \leqslant y-1$ and $1 \leqslant y \leqslant c$, the sub-matrix $\mathbf{S}(y; y')$ has dimension $(y+1) \times (y'+1)$, and its elements are given by

$$(\mathbf{S}(y; y'))_{k,k'} = \begin{cases} \nu^{y-y'} \frac{(k-1)!}{(k'-1)!}, & \text{if } y - y' + 1 \leqslant k \leqslant y+1 \quad \text{and} \quad k' = k - y + y', \\ 0, & \text{otherwise.} \end{cases}$$

(b) For $0 \leqslant y' \leqslant c-1$ and $c+1 \leqslant y \leqslant 2c$, the sub-matrix $\mathbf{S}(y; y')$ has dimension $(c+1) \times (y'+1)$, and its elements are given by

$$(\mathbf{S}(y; y'))_{k,k'} = \begin{cases} \nu^{y-y'} \frac{(k-1)!}{(k'-1)!}, & \text{if } y - y' + 1 \leqslant k \leqslant c+1 \quad \text{and} \quad k' = k - y + y', \\ 0, & \text{otherwise.} \end{cases}$$

In particular, $\mathbf{S}(y; y') = \mathbf{0}_{(c+1)\times(y'+1)}$ if $y - y' > c$.

(c) For $0 \leqslant y' \leqslant c-1$ and $2c+1 \leqslant y$, we have $\mathbf{S}(y; y') = \mathbf{0}_{(c+1)\times(y'+1)}$.

(d) For $c \leqslant y' < y - c$ and $c+1 \leqslant y$, we have $\mathbf{S}(y; y') = \mathbf{0}_{(c+1)\times(c+1)}$.

(e) For $y - c \leqslant y' \leqslant y-1$ and $c+1 \leqslant y$, the sub-matrix $\mathbf{S}(y; y')$ is a square matrix of order $c+1$, and it consists of the elements

$$(\mathbf{S}(y; y'))_{k,k'} = \begin{cases} \nu^{y-y'} \frac{(k-1)!}{(k'-1)!}, & \text{if } y - y' + 1 \leqslant k \leqslant c+1 \quad \text{and} \quad k' = k - y + y', \\ 0, & \text{otherwise.} \end{cases}$$

References

[1] S. Aguir, F. Karaesmen, O.Z. Akşin, F. Chauvet, The impact of retrials on call center performance, OR Spectr. 26 (2004) 353–376.
[2] J.R. Artalejo, M. Pozo, Numerical calculation of the stationary distribution of the main multiserver retrial queue, Ann. Oper. Res. 116 (2002) 41–56.
[3] J.R. Artalejo, A. Economou, A. Gómez-Corral, Applications of maximum queue lengths to call center management, Comput. Oper. Res. 34 (2007) 983–996.
[4] J.R. Artalejo, A. Economou, M.J. Lopez-Herrero, Algorithmic analysis of the maximum queue length in a busy period for the $M/M/c$ retrial queue, INFORMS J. Comput. 19 (2007) 121–126.
[5] J.R. Artalejo, A. Gómez-Corral, Retrial Queueing Systems. A Computational Approach, Springer-Verlag, Berlin, 2008.
[6] J.R. Artalejo, On the transient behavior of the maximum level length in structured Markov chains, in: L. Pardo, N. Balakrishnan, M.A. Gil (Eds.), Modern Mathematical Tools and Techniques in Capturing Complexity, Springer Series in Synergetics, Berlin, 2011, pp. 332–342.
[7] L. Bright, P.G. Taylor, Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes, Stoch. Models 11 (1995) 497–525.
[8] J.W. Cohen, Basic problems of telephone traffic theory and the influence of repeated calls, Philips Telecomm. Rev. 18 (1957) 49–100.
[9] T. Dayar, Analyzing Markov Chains Using Kronecker Products: Theory and Applications, Springer Briefs in Mathematics, Springer, New York, 2012.
[10] G.I. Falin, Calculation of probability characteristics of a multiline system with repeat calls, Mosc. Univ. Comput. Math. Cybern. 1 (1983) 43–49.
[11] G.I. Falin, J.G.C. Templeton, Retrial Queues, Chapman & Hall, London, 1997.
[12] A.A. Fredericks, G.A. Reisner, Approximations to stochastic service systems, with an application to a retrial model, Bell Syst. Tech. J. 58 (1979) 557–576.
[13] A. Gómez-Corral, M.F. Ramalhoto, The stationary distribution of a Markovian process arising in the theory of multiserver retrial queueing systems, Math. Comput. Model. 30 (1999) 141–158.
[14] A. Gómez-Corral, M. López García, Maximum population sizes in host-parasitoid models, Int. J. Biomath. 6 (2) (2013) 1350002, http://dx.doi.org/10.1142/S1793524513500022. 28 pages.
[15] B.S. Greenberg, R.W. Wolff, An upper bound on the performance of queues with returning customers, J. Appl. Probab. 24 (1987) 466–475.
[16] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1985.
[17] J.J. Hunter, Mathematical Techniques of Applied Probability, Discrete Time Models: Basic Theory, vol. 1, Academic Press, New York, 1983.
[18] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, ASA-SIAM Series on Statistics and Applied Probability, Philadelphia, 1999.

[19] D.M.B. Masi, M.J. Fischer, C.M. Harris, Computation of steady-state probabilities for resource-sharing call-center queueing systems, Stoch. Models 17 (2001) 191–214.
[20] C. Moler, C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later, SIAM Rev. 45 (2003) 3–49.
[21] M.F. Neuts, B.M. Rao, Numerical investigation of a multiserver retrial model, Queueing Syst. 7 (1990) 169–189.
[22] C.E.M. Pearce, Extended continued fractions, recurrence relations and two-dimensional Markov processes, Adv. Appl. Probab. 21 (1989) 357–375.
[23] T. Phung-Duc, H. Masuyama, S. Kasahara, Y. Takahashi, $M/M/3/3$ and $M/M/4/4$ retrial queues, J. Ind. Manage. Optim. 5 (2009) 431–451.
[24] T. Phung-Duc, H. Masuyama, S. Kasahara, Y. Takahashi, State-dependent $M/M/c/c+r$ retrial queues with Bernoulli abandonment, J. Ind. Manage. Optim. 6 (2010) 517–540.
[25] T. Phung-Duc, H. Masuyama, S. Kasahara, Y. Takahashi, A simple algorithm for the rate matrices of level-dependent QBD processes, in: Proceedings of QTNA'10, Fifth International Conference on Queueing Theory and Network Applications, ACM Digital Library, New York, 2010, pp. 46–52.
[26] T. Phung-Duc, H. Masuyama, S. Kasahara, Y. Takahashi, A matrix continued fraction approach to multiserver retrial queues, Ann. Oper. Res. 202 (2013) 161–183.
[27] J. Riordan, Stochastic Service Systems, Wiley, New York, 1962.
[28] S.N. Stepanov, Markov models with retrials: the calculation of stationary performance measures based on the concept of truncation, Math. Comput. Model. 30 (1999) 207–228.
[29] T. Strom, On logarithmic norms, SIAM J. Numer. Anal. 12 (1975) 741–753.
[30] H.F. Trotter, On the product of semigroups of operators, Proc. Am. Math. Soc. 10 (1959) 545–551.
[31] R.I. Wilkinson, Theories for toll traffic engineering in the U.S.A, Bell Syst. Tech. J. 35 (1956) 421–514.
[32] R.W. Wolff, Stochastic Modeling and the Theory of Queues, Prentice-Hall, New Jersey, 1989.