The Sixth International Symposium on Operations Research and Its Applications (ISORA'06)
Xinjiang, China, August 8–12, 2006
Copyright © 2006 ORSC & APORC, pp. 60–74

An Algorithmic Solution for the Stationary Distribution of $M/M/c/K$ Retrial Queue

Yang Woo Shin[1,*]   Dug Hee Moon[2,†]

[1] Department of Statistics, Changwon National University, Changwon, Gyeongnam 641-773, Korea
[2] Department of Industrial and Systems Engineering, Changwon National University, Changwon, Gyeongnam 641-773, Korea

Abstract We present an algorithmic solution for the stationary distribution of $M/M/c/K$ retrial queue which consists of an orbit with infinite capacity and a service facility that has c exponential servers and waiting space of size $K-c$. The behavior of queue length process in the retrial queue is described by level dependent quasi-birth-and-death (LDQBD) process due to repeated attempts.

The algorithm is based on the generalized truncation method (GTM) proposed by Nuets and Rao [9] which is use of the level independent QBD process except the first N levels as an approximation of the original LDQBD process and the truncation level N is enlarged until the satisfactory solution is obtained. As the authors indicated, the method in Nuets and Rao [9] may not perform very well when the system is highly congested. Main features of our algorithm are to develop a very simple and effective method for deriving inverse of the matrices in the diagonal blocks and to provide the stable and efficient ways for computing the rate matrices and the boundary probability vectors. Our approach can overcome drawbacks of the algorithms in Nuets and Rao [9] and can be applied not only to the system with very large number s of servers and very large size $K-s$ of waiting space but also to the highly congested system.

Keywords multi-server retrial queue; finite buffer; level dependent quasi-birth-and-death (LDQBD) process; generalized truncation method; matrix geometric solutions

1 Introduction

We consider the $M/M/c/K$ retrial queue which consists of an orbit with infinite capacity and a service facility that has c exponential servers and waiting space of size $K-c$. Retrial queues are characterized by the following features. When an arriving customer finds that all servers are busy and no waiting position is available, the customer joins a virtual pool of blocked customers called orbit and repeats its request after a random amount of time, called retrial time until the customer gets into

*This research was supported by the MIC(Ministrary of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment). E-mail: ywshin@changwon.ac.kr

†This work is partially supported by the Ministry of Commerce, Industry and Energy(MOCIE) through the fostering project of the Regional Industry. E-mail: dhmoon@changwon.ac.kr

the service facility. We assume that the access from orbit to the service facility is governed by the exponential distribution whose rate may depend on the current number of customers in orbit. The behavior of queue length process in the retrial queue is described by level dependent quasi-birth-and-death (LDQBD) process due to repeated attempts. For general treatment of the calculation of stationary distribution of LDQBD process, see Bright and Taylor [4]. This spatial inhomogeneity often leads the analytical complexity and approximations are used instead. For an approximation of the stationary distribution of the retrial queueing systems, many authors use another calculable system with infinite state space, so-called generalized truncated system that is a Markov chain with spatially homogeneous block-partitioned generator except the first N levels e.g., see [1, 2, 5, 9, 12] and the references therein. The detailed overviews of the related references with retrial queues can be found in [3, 5].

In this paper we propose a numerical algorithm for computing the stationary distribution. The algorithms are based on the generalized truncation method (GTM) proposed by Nuets and Rao [9] which uses of the level independent QBD process except first N levels as an approximation of the original LDQBD process and the truncation level N is enlarged until the satisfactory solution is obtained. Shin [11] shows that the stationary distribution of the generalized truncated system in many retrial queues including $M/M/c/K$ retrial queue converges to that of the original system as the truncation level N becomes infinity. The GTM consists of the following three steps. The first step is to modify the infinitesimal generator, say Q of the original LDQBD process to Q_N, the generator of the level independent QBD process except first N levels. The second step is to find the stationary distribution $\boldsymbol{y}$ of Q_N and to increase N until the individual elements of $\boldsymbol{y}$ do not change significantly. Finally, approximate the stationary distribution $\boldsymbol{x}$ of Q by $\boldsymbol{y}$. For $\boldsymbol{y}$ corresponding to the level of the homogeneous part of Q_N, matrix geometric method in Neuts [8] are used and the probabilities corresponding to the boundary levels are obtained by are obtained by solving the system of linear equations. In Neuts and Rao [9], the successive substitution method in Neuts [8] is used for rate matrix and the block Gauss-Seidel scheme is used for the boundary probabilities. Thus as the truncation level N increases, the size of linear system for boundary probabilities becomes large. This situation can occur when the system is highly congested. Furthermore as the size K of service facility increases, the size of block matrix components of Q_N also increases. Thus the method in Neuts and Rao [9] may not perform very well under the conditions of severe congestion when the traffic intensity is high and retrial rate is very small as the authors indicated. To overcome drawbacks in Neuts and Rao [9], we develop a simple and effective method for deriving the inverse of the matrices in diagonal blocks and provide the stable and efficient ways for computing the rate matrices and the boundary probability vectors. Our approach can be applied not only to the system with very large size K of service facility but also to the highly congested system.

This paper is organized as follows. In section 2, the mathematical model is described in detail. The algorithmic solutions are proposed in section 3. Section 4 deals with numerical results and concluding remarks are presented in section 5.

2 Model description

Consider an $M/M/c/K$ retrial queue which consists of infinite capacity of orbit and the service facility with c identical servers and $K-c$ waiting positions. Service times of customers are independent of each other and have a common exponential distribution with parameter μ. Customers arrive according to a Poisson process with rate λ. The customer who finds that all servers are busy and no waiting position is available upon its arrival joins orbit and tries to its luck again after a random amount of time until the customer gets into the service facility. The access from orbit to the service facility is governed by the exponential distribution with rate γ_k which may depend on the current number k, $k \geq 0$ of customers in orbit. That is, the probability of repeated attempt during the interval $(t,\, t+\Delta t)$, given that k customers in orbit at time t, is $\gamma_k \Delta t + o(\Delta t)$ with $\gamma_0 = 0$. We assume that the retrial rates γ_k, $k \geq 0$ satisfy $\gamma_k \leq \gamma_{k+1}$ ($\gamma_0 = 0$) and $\lim_{k\to\infty} \gamma_k = \infty$. Let $X_0(t)$ and $X_1(t)$ be the number of customers in orbit and in the service facility at time t, respectively and $\boldsymbol{X} = \{X(t), t \geq 0\}$ with $X(t) = (X_0(t), X_1(t))$. Then the stochastic process $\boldsymbol{X}$ is a Markov chain on the state space $\mathscr{S} = \{(i,j) : i = 0,1,2,\cdots,\ j = 0,1,\cdots,K\}$. Setting $\mathbf{k} = \{(k,0),(k,1),\cdots,(k,K)\}$, $k \geq 0$, the generator of the Markov chain $\boldsymbol{X}$ is of the form

$$Q = \begin{array}{c} \\ \mathbf{0} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \cdots \end{array} \\ \begin{pmatrix} A_1^{(0)} & A_0 & & & \\ A_2^{(1)} & A_1^{(1)} & A_0 & & \\ & A_2^{(2)} & A_1^{(2)} & A_0 & \\ & & A_2^{(3)} & A_1^{(3)} & \cdots \\ & & & \vdots & \end{pmatrix} \end{array}, \tag{1}$$

where A_0, $A_1^{(n)}$ and $A_2^{(n)}$ are square matrices of order $K+1$ and and are given by

$$\mathbf{A_0} = \begin{pmatrix} & O & \\ & & \\ & & \lambda \end{pmatrix}, \quad A_2^{(n)} = \gamma_n \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}, \; n \geq 1,$$

and

$$A_1^{(n)} = \begin{pmatrix} -\Lambda_{n,0} & \lambda & & & & & & \\ \mu & -\Lambda_{n,1} & \lambda & & & & & \\ & 2\mu & -\Lambda_{n,2} & \lambda & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & s\mu & -\Lambda_{n,c} & \lambda & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & s\mu & -\Lambda_{n,c} & \lambda \\ & & & & & & s\mu & -\Lambda_{0,c} \end{pmatrix}, \; n \geq 0,$$

with $\Lambda_{n,k} = \lambda + \gamma_n + k\mu$, $n \geq 0$.

We assume that $\rho = \frac{\lambda}{c\mu} < 1$ which guarantees the existence of the stationary distribution of $\boldsymbol{X}$ (see He et al. [6] and Shin [10]). Let $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots)$ with $\boldsymbol{x}_n = (x_{n0}, \cdots, x_{nK})$, $n \geq 0$ be the stationary distribution of Q. In order to calculate the stationary distribution $\boldsymbol{x}$, it is necessary to have the family of square matrices $\{R_k, k \geq 0\}$ of order $K+1$ which are the minimal nonnegative solutions to the systems of equations

$$A_0 + R_k A_1^{(k+1)} + R_k(R_{k+1} A_2^{(k+2)}) = 0, \ k \geq 0. \tag{2}$$

It follows from the special structure of the matrix A_0 that

$$R_k = A_0(-A_1^{(k+1)} - R_{k+1} A_2^{(k+2)})^{-1} \tag{3}$$

has the following formula

$$R_k = \begin{pmatrix} O \\ \boldsymbol{r}_k \end{pmatrix}, \ k \geq 0, \tag{4}$$

where O is the zero matrix of size $K \times (K+1)$ and $\boldsymbol{r}_k = (r_{k0}, r_{k1}, \cdots, r_{kK})$ is the $(K+1)-$ row vector. Thus it follows from Bright and Taylor [4] that the stationary distribution $\boldsymbol{x} = (\boldsymbol{x}_i, i \geq 0)$ is given by

$$\boldsymbol{x}_n = \boldsymbol{x}_0 \left(\prod_{k=0}^{n-1} R_k \right) = x_{0K} \left(\prod_{k=0}^{n-2} r_{kK} \right) \boldsymbol{r}_{n-1}, \ n \geq 1, \tag{5}$$

and $\boldsymbol{x}_0$ is the unique solution of the equation

$$\boldsymbol{x}_0(A_1^{(0)} + R_0 A_2^{(1)}) = 0 \tag{6}$$

with the normalizing condition

$$\boldsymbol{x}_0 \mathbf{1} + x_{0K} \left(\boldsymbol{r}_0 \mathbf{1} + \sum_{n=2}^{\infty} \left(\prod_{k=0}^{n-2} r_{kK} \right) \boldsymbol{r}_{n-1} \mathbf{1} \right) = 1, \tag{7}$$

where $\mathbf{1}$ is the column vector of corresponding size whose components are all 1.

3 Algorithmic Solution

In this section, we present an algorithmic solution for the stationary distribution $\boldsymbol{x}$ of Q by using the generalized truncation method. The first step of the approximation is to modify the infinitesimal generator Q to Q_N by letting $A_i^{(k)} = A_i^{(N)}$, $i = 1, 2$ for

$k \geq N$, where N is a fixed positive integer. That is, Q_N is of the form

$$Q_N = \begin{array}{c} \\ \mathbf{0} \\ \mathbf{1} \\ \vdots \\ \mathbf{N-1} \\ \mathbf{N} \\ \mathbf{N+1} \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccccc} \mathbf{0} & \mathbf{1} & \cdots & \mathbf{N-1} & \mathbf{N} & \mathbf{N+1} & \cdots & \end{array} \\ \begin{pmatrix} A_1^{(0)} & A_0 & & & & & \\ A_2^{(1)} & A_1^{(1)} & A_0 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & A_2^{(N-1)} & A_1^{(N-1)} & A_0 & & \\ & & & A_2 & A_1 & A_0 & \\ & & & & A_2 & A_1 & \cdots \\ & & & & & \vdots & \end{pmatrix} \end{array}, \tag{8}$$

where $A_1 = A_1^{(N)}$ and $A_2 = A_2^{(N)}$. It is well known (e.g. see Neuts [8]) that Q_N is positive recurrent if and only if

$$\rho(N) = \frac{\boldsymbol{\pi}^{(N)} A_0 \mathbf{1}}{\boldsymbol{\pi}^{(N)} A_2^{(N)} \mathbf{1}} = \frac{\lambda \pi_K^{(N)}}{\gamma_N \left(1 - \pi_K^{(N)}\right)} < 1, \tag{9}$$

where $\boldsymbol{\pi}^{(N)} = (\pi_j^{(N)}, 0 \leq j \leq K)$ is the stationary distribution of $A^{(N)} = A_0 + A_1^{(N)} + A_2^{(N)}$, that is, $\boldsymbol{\pi}^{(N)} A^{(N)} = \mathbf{0}$ and $\boldsymbol{\pi}^{(N)} \mathbf{1} = 1$. Noting that $A^{(N)}$ is the same as the generator of the Markov chain for queue length process in the ordinary $M/M/c/K$ queue with arrival rate $\lambda + \gamma_N$ and service rate μ of each server, it is easily seen that $\pi_K^{(N)}$ is given by

$$\pi_K^{(N)} = \frac{c^c}{c!} (\hat{\rho}_N)^K \left[\sum_{j=0}^{c-1} \frac{(c\hat{\rho}_N)^j}{j!} + \frac{(c\hat{\rho}_N)^c}{c!} \frac{1 - (\hat{\rho}_N)^{K-c+1}}{1 - \hat{\rho}_N} \right]^{-1},$$

where $\hat{\rho}_N = \frac{\lambda + \gamma_N}{c\mu}$. By showing that $\lim_{N \to \infty} \boldsymbol{\pi}^{(N)} = \rho$, it can be seen that the Markov chain Q_N is positive recurrent under the condition $\rho < 1$ for sufficiently large N. A sufficient condition $\rho < 1$ of ergodicity of Q_N for sufficiently large N can also be obtained from Shin [10]. The second step is to find the stationary distribution $\boldsymbol{y}$ of Q_N and to increase N until the individual elements of $\boldsymbol{y}$ do not change significantly. Note that $\boldsymbol{y}$ depends on the truncation level N. We shall write $\boldsymbol{y}(N)$ instead of $\boldsymbol{y}$, whenever necessary for clarity. Finally, approximate $\boldsymbol{x}$ by $\boldsymbol{y}$.

Let S is the minimal nonnegative solution of the matrix equation

$$A_0 + SA_1 + S^2 A_2 = 0 \tag{10}$$

and

$$S_k = A_0(-A_1^{(k+1)} - S_{k+1} A_2^{(k+2)})^{-1},\ k = N-2, N-3, \cdots, 1, 0 \tag{11}$$

with $S_{N-1} = S$. It follows from the special structure of the matrix A_0 that S and S_k, $0 \leq k \leq N-2$ have the following formulae

$$S = \begin{pmatrix} O \\ \boldsymbol{s} \end{pmatrix}, \quad S_k = \begin{pmatrix} O \\ \boldsymbol{s}_k \end{pmatrix},\ 0 \leq k \leq N-2, \tag{12}$$

where O is the zero matrix of size $K \times (K+1)$, and $\boldsymbol{s} = (s_0, s_1, \cdots, s_K)$ and $\boldsymbol{s}_k = (s_{k0}, s_{k1}, \cdots, s_{kK})$ are the $K+1$ row vector. To highlight the matrices S is a function of N, we shall write $S(N)$ and $\boldsymbol{s}(N)$ instead of S and $\boldsymbol{s}$, whenever necessary for clarity. Similarly, $S_k(N)$ and $\boldsymbol{s}_k(N)$ will be used instead of S_k and $\boldsymbol{s}_k$.

Write $\boldsymbol{y} = (\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots)$ in the block partitioned form, where $\boldsymbol{y}_i = (y_{i0}, \cdots, y_{iK})$, $i \geq 0$. Following the similar procedures to those of (5), (6) and (7), $\boldsymbol{y}_i$, $i \geq 0$ are given by

$$\boldsymbol{y}_n = \begin{cases} y_{0K}\left(\prod_{k=0}^{n-2} s_{kK}\right)\boldsymbol{s}_{n-1}, & 1 \leq n \leq N-1, \\ y_{0K}\left(\prod_{k=0}^{N-2} s_{kK}\right)(s_K)^{n-N}\boldsymbol{s}, & n \geq N. \end{cases} \tag{13}$$

The vector $\boldsymbol{y}_0$ is the unique solution of the equation

$$\boldsymbol{y}_0(A_1^{(0)} + S_0 A_2^{(1)}) = 0 \tag{14}$$

with the normalizing condition

$$\boldsymbol{y}_0\mathbf{1} + y_{0K}\left[\boldsymbol{s}_0\mathbf{1} + \sum_{n=2}^{N-1}\left(\prod_{k=0}^{n-2} s_{kK}\right)\boldsymbol{s}_{n-1}\mathbf{1} + \left(\prod_{k=0}^{N-2} s_{kK}\right)\frac{\boldsymbol{s}\mathbf{1}}{1-s_K}\right] = 1. \tag{15}$$

So the algorithm consists of two parts, (1) computation of S and S_k, and (2) choice of truncation level N.

Computations of S and S_k. Since $S^2 = s_K S$, we have from (10) and (12) that

$$\boldsymbol{s} = (\lambda \mathbf{e}_{K+1}^T + \gamma_N s_K \tilde{\boldsymbol{s}})(-A_1)^{-1}, \tag{16}$$

where $\mathbf{e}_{K+1}^T = (0, \cdots, 0, 1)$ is a $(K+1)$-dimensional row vector and $\tilde{\boldsymbol{s}} = (0, s_0, s_1, \cdots, s_{K-1})$. That is, the jth entry s_j of $\boldsymbol{s}$ is given by

$$s_j = \lambda a_{K+1,j+1} + \gamma_N s_K \sum_{i=1}^{K} s_{i-1} a_{i+1,j+1},\ j = 0, 1, \cdots, K, \tag{17}$$

where $(-A_1)^{-1} = (a_{ij})_{1 \leq i,j \leq K+1}$. The solution of the equation (16) is obtained by $\boldsymbol{s} = \lim_{n\to\infty} \boldsymbol{s}^{(n)}$, where the sequence $\{\boldsymbol{s}^{(n)}, n \geq 0\}$ is defined by

$$\boldsymbol{s}^{(n+1)} = (\lambda \mathbf{e}_{K+1}^T + \gamma_N s_K^{(n)} \tilde{\boldsymbol{s}}^{(n)})(-A_1)^{-1} \tag{18}$$

with $\boldsymbol{s}^{(0)} = \mathbf{0}$ and $\tilde{\boldsymbol{s}}^{(n)} = (0, s_0^{(n)}, s_1^{(n)}, \cdots, s_{K-1}^{(n)})$. Since all the elements in (18) are nonnegative, the iteration (18) provides a stable method for evaluating S. To expedite the convergence to $\boldsymbol{s}$, we use a modified method that uses the new one $s_i^{(n+1)}$ for $i < j$ for computing $s_j^{(n+1)}$ as

$$s_j^{(n+1)} = \lambda a_{K+1,j+1} + \gamma_N s_K^{(n)}\left(\sum_{i=1}^{j-1} s_{i-1}^{(n+1)} a_{i+1,j+1} + \sum_{i=j}^{K} s_{i-1}^{(n)} a_{i+1,j+1}\right),\ j = 0, 1, \cdots, K. \tag{19}$$

Iteration (19) is repeated until for arbitrarily chosen $\varepsilon_0 > 0$, the tolerance condition

$$||\boldsymbol{s}^{(n+1)} - \boldsymbol{s}^{(n)}||_\infty = \max_{0\le j\le K} |s_j^{(n+1)} - s_j^{(n)}| < \varepsilon_0 \tag{20}$$

is satisfied.

We can see from (16) and (11) that it is necessary to evaluate $(-A_1)^{-1}$ and $(-A_1^{(k+1)} - S_{k+1}A_2^{(k+2)})^{-1}$, $0 \le k \le N-1$ for computing S and S_k. The following results for $(-A_1^{(k)})^{-1}$ are immediate from Proposition A.2 in Appendix. Let $\mu_j = \min(j,c)\mu$, $0 \le j \le K+1$. Define $w_j^{(k)}$ and $v_j^{(k)}$, $1 \le j \le K+1$ as follows:

$$\begin{aligned}
w_{K+1}^{(k)} &= \frac{\lambda}{\lambda + c\mu},\\
w_j^{(k)} &= \frac{\lambda}{\lambda + \gamma_k + \mu_{j-1} - w_{j+1}^{(k)}\mu_j},\ j = K, K-1, \cdots, 2,\\
w_1^{(k)} &= \frac{1}{\lambda + \gamma_k - w_2^{(k)}\mu}
\end{aligned}$$

and

$$\begin{aligned}
v_{K+1}^{(k)} &= \frac{c\mu}{\lambda + c\mu},\\
v_j^{(k)} &= \frac{\mu_{j-1}}{\lambda(1 - v_{j+1}^{(k)}) + \gamma_k + \mu_{j-1}},\ j = K, K-1, \cdots, 2,\\
v_1^{(k)} &= \frac{1}{\lambda(1 - v_2^{(k)}) + \gamma_k}.
\end{aligned}$$

Proposition 1. *The inverse matrix $(-A_1^{(k)})^{-1} = (a_{ij}^{(k)})_{1\le i,j\le K+1}$ of $-A_1^{(k)}$ is given as follows: (1) The first row and column :*

$$a_{1j}^{(k)} = w_1^{(k)} w_2^{(k)} \cdots w_j^{(k)},\ j = 1, 2, \cdots, K+1, \tag{21}$$

$$a_{i1}^{(k)} = v_i^{(k)} v_{i-1}^{(k)} \cdots v_1^{(k)},\ i = 1, 2, \cdots, K+1, \tag{22}$$

(2) The j-th $(2 \le j \le K)$ row and column :

$$a_{jj}^{(k)} = \frac{1 + \lambda a_{j,j-1}^{(k)}}{\lambda + \gamma_k + \mu_{j-1} - w_{j+1}^{(k)}\mu_j}, \tag{23}$$

$$a_{jm}^{(k)} = a_{jj}^{(k)} w_{j+1}^{(k)} w_{j+2}^{(k)} \cdots w_m^{(k)},\ m = j+1, j+2, \cdots, K+1, \tag{24}$$

$$a_{mj}^{(k)} = v_m^{(k)} v_{m-1}^{(k)} \cdots v_{j+1}^{(k)} a_{jj}^{(k)},\ m = j+1, j+2, \cdots, K+1, \tag{25}$$

(3) The $(K+1, K+1)$-component :

$$a_{K+1,K+1}^{(k)} = \frac{1 + \lambda a_{K+1,K}^{(k)}}{\lambda + c\mu}. \tag{26}$$

For the inverse $(-A_1^{(k+1)} - S_{k+1}A_2^{(k+2)})^{-1}$, we need the following lemma.

Lemma 2. *Let A be an invertible matrix of size n and*

$$B = A + a\boldsymbol{u}\boldsymbol{v}^T,$$

where $\boldsymbol{u}$ and $\mathbf{v}$ are column n-vectors and $\mathbf{v}^T$ is the transpose of $\mathbf{v}$. If B is invertible, then the inverse matrix B^{-1} of B is given by

$$B^{-1} = A^{-1} - \frac{a}{1 + a\boldsymbol{v}^T A^{-1}\boldsymbol{u}} A^{-1}\boldsymbol{u}\boldsymbol{v}^T A^{-1}.$$

In particular, if $\boldsymbol{u} = \boldsymbol{v} = \boldsymbol{e}_k = (0,\cdots,0,1,0,\cdots,0)^T$ is the column n-vector whose kth component is 1 and others are all zeros, then the (i,j) component $[B^{-1}]_{ij}$ of B^{-1} is

$$[B^{-1}]_{ij} = [A^{-1}]_{ij} - \frac{a}{1 + a[A^{-1}]_{kk}}[A^{-1}]_{ik}[A^{-1}]_{kj},\ 1 \le i,j \le n.$$

Proposition 3. *Once $\boldsymbol{s}_{k+1}$ is obtained, $\boldsymbol{s}_k = (s_{k0}, s_{k1}, \cdots, s_{kK})$ is given by*

$$s_{kj} = \lambda a_{K+1,j+1}^{(k+1)} + \lambda \frac{\left(\gamma_{k+2} a_{K+1,K+1}^{(k+1)}\right)\left(\sum_{i=1}^{K} s_{k+1,i-1} a_{i+1,j+1}^{(k+1)}\right)}{1 - \gamma_{k+2}\left(\sum_{i=1}^{K} s_{k+1,i-1} a_{i+1,K+1}^{(k+1)}\right)},\ j = 0,1,2,\cdots,K. \quad (27)$$

Proof. Noting that

$$A_1^{(k+1)} + S_{k+1}A^{(k+2)} = A_1^{(k+1)} + \gamma_{k+2}\mathbf{e}_{K+1}\tilde{\boldsymbol{s}}_{k+1},$$

(27) is immediate from Lemma 2. □

Choice of N. The choice of the truncation level N is important because the amount of computation and the approximation error depend on it critically. It can be seen from (13) that the tail distribution of $\boldsymbol{y}$ depends largely on the spectral radius $sp(S(N)) = s_K(N)$ of $S(N)$. As Neuts and Rao [9] indicated, to minimize the effect of the approximation, the truncation level N must be chosen such that $s_K(N)$ is sufficiently close to $s_K(\infty)$. In the following, we show that $sp(R_N) = r_{N,K}$ and $sp(S(N)) = s_K(N)$ have the same limit ρ as N tends to infinity.

Lemma 4.

$$\lim_{k\to\infty} r_{kK} = \rho = \lim_{N\to\infty} s_K(N). \quad (28)$$

Proof. Let $\Delta_k = -diag[A_1^{(k)}]$ be the diagonal matrix whose diagonal elements are $[-A_1^{(k)}]_{ii}$, $1 \le i \le K+1$ and define

$$\begin{aligned} P_0^{(k)} &= \Delta_k^{-1}A_0,\ k \ge 0, \\ P_1^{(k)} &= \Delta_k^{-1}A_1^{(k)} + I,\ k \ge 0, \\ P_2^{(k)} &= \Delta_k^{-1}A_2^{(k)},\ k \ge 1 \end{aligned}$$

and let

$$R_k^J = \Delta_k^{-1} R_k \Delta_{k+1},\ k \geq 0. \tag{29}$$

Then the matrices R_k^J, $k \geq 0$ are the minimal nonnegative solutions to the equations

$$R_k^J = P_0^{(k)} + R_k^J P_1^{(k+1)} + R_k^J [R_{k+1}^J P_2^{(k+2)}],\ k \geq 0. \tag{30}$$

We have from (29) that R_k^J has non-zero elements only in the last row and denote it by $\boldsymbol{r}_k^J = (r_{k0}^J, \cdots, r_{k,K}^J)$ and

$$\boldsymbol{r}_k = \left(\frac{\lambda + c\mu}{\Lambda_{k+1,0}} r_{k,0}^J, \frac{\lambda + c\mu}{\Lambda_{k+1,1}} r_{k,1}^J, \cdots, \frac{\lambda + c\mu}{\Lambda_{k+1,K-1}} r_{k,K-1}^J, r_{k,K}^J \right).$$

Thus $r_{k,K} = r_{k,K}^J$ and hence $r_{\infty,K} = r_{\infty,K}^J$. Letting $k \to \infty$ in (30), we have that

$$\begin{aligned}
&r_{\infty,0}^J = 0,\\
&r_{\infty,j}^J = r_{\infty,K}^J r_{\infty,j-1}^J,\ j = 1, 2, \cdots, K-2,\\
&r_{\infty,K-1}^J = r_{\infty,K}^J \frac{c\mu}{\lambda + c\mu} + r_{\infty,K}^J r_{\infty,K-2}^J,\\
&r_{\infty,K}^J = \frac{\lambda}{\lambda + c\mu} + r_{\infty,K}^J r_{\infty,K-1}^J.
\end{aligned}$$

The nonnegative solutions that are less than 1 of the equations above are given as follows:

$$\begin{aligned}
&r_{\infty,j}^J = 0,\ j = 0, 1, \cdots, K-2,\\
&r_{\infty,K-2}^J = \frac{\rho}{1+\rho},\\
&r_{\infty,K}^J = \rho.
\end{aligned}$$

Thus $r_{\infty,K} = r_{\infty,K}^J = \rho$ and hence the first part of (28) is proved. We can show the second equality in (28) by noting that

$$S(N) = \Delta_N S^J(N) \Delta_N^{-1}$$

and $S^J(N)$ is a minimal nonnegative solution of the equation

$$S^J(N) = P_0^{(N)} + S^J(N) P_1^{(N)} + (S^J(N))^2 P_2^{(N)}.$$

The details are omitted. □

Remark 1. Neuts and Rao [9] claim that $\lim_{N\to\infty} sp(S(N)) = \rho$ by numerical experiments for various case. The model considered in [9] is a little bit different from that in this paper. But our method can be applied to the case of Neuts and Rao [9].

Now we propose criteria for selecting N. Basically, N must be large enough for Q_N to satisfy the stability condition (9) and the stationary distribution $\boldsymbol{y}$ exists. We choose N which satisfies the following conditions for $S(N)$:

$$|s_K(N)-\rho| < \varepsilon_1 \text{ or } ||S(N)-S(N-1)||_\infty < \varepsilon_2 \tag{31}$$

for arbitrarily given $\varepsilon_i > 0$, $i = 1,2$. Furthermore, to reduce the effect of approximation, we choose to increase N until the tail distribution $\sum_{k=N}^{\infty} \boldsymbol{y}_k(N)\mathbf{1} < \varepsilon_3$ for arbitrarily given $\varepsilon_3 > 0$. The following algorithm summarizes the results above.

Algorithm 1. Approximation for $\boldsymbol{x}$

1. Choose an initial level N_0 for $\rho(N_0) < 1$ and let $N := N_0$.
2. Compute $(-A_1^{(N)})^{-1}$ and $S(N)$ using Algorithm A and (19), respectively.
3. **IF** $(|s_K(N)-\rho| < \varepsilon_1$ or $||S(N)-S(N-1)||_\infty < \varepsilon_2)$ **THEN**
   compute $S_k(N)$, $k = N-1,\cdots,0$ and $\boldsymbol{y}$ using (27) and (13), respectively;
   ELSE $N := N+1$ and GO TO 2. **ENDIF**;
4. Check the criterion for tail distribution of $\boldsymbol{y}(N)$.
   IF $(\sum_{k=N}^{\infty} \boldsymbol{y}_k(N)\mathbf{1} < \varepsilon_3)$ **THEN**
   STOP;
   ELSE $N := N+1$ and GO TO 2. **ENDIF**;

4 Numerical results

We consider $M/M/25/K$ retrial queue with arrival rate $\lambda = 1.0$ and retrial rate $\gamma_k = k\gamma$, $k \geq 0$. To show the effectiveness and feasibility of algorithm 1 to the case of large K and/or highly congested system, where ρ is high and γ is very small comparing the arrival rate λ, we consider the following combinations $\rho = 0.9$, 0.95, 0.975, $\gamma = 10.0$, 1.0, 0.5,0.1, 0.05, $c = 25$, and $K = 25$, $K = 50$. Some system characteristics such as the mean number L_0 of customers in orbit and the mean number L_1 of customers in service facility, and the blocking probability BP are considered. The expressions for L_0, L_1 and BP are as follows:

$$\begin{aligned}
L_0 &= \left(\sum_{k=0}^{\infty} k\boldsymbol{y}_k\right)\mathbf{1} = \left(\sum_{k=0}^{N-1} k\boldsymbol{y}_k\right)\mathbf{1} + y_{N-1,K}\left(\frac{s_K}{(1-s_K)^2} + \frac{N}{1-s_K}\right)\boldsymbol{s}\mathbf{1},\\
L_1 &= \left(\sum_{k=0}^{\infty} \boldsymbol{y}_k\right)\boldsymbol{l}_1 = \left(\sum_{k=0}^{N-1} \boldsymbol{y}_k + y_{N-1,K}\frac{\boldsymbol{s}}{1-s_K}\right)\boldsymbol{l}_1,\\
BP &= \sum_{n=0}^{\infty} \boldsymbol{y}_{n,K} = \sum_{n=0}^{N-1} \boldsymbol{y}_{n,K} + y_{N-1,K}\frac{s_K}{1-s_K},
\end{aligned}$$

where $\boldsymbol{l}_1 = (0,1,2,\cdots,K)^T$. We choose the stopping criteria $\varepsilon_0 = 0.001$ for (20), $\varepsilon_1 = \varepsilon_2 = 0.001$ and $\varepsilon_3 = 0.0001$ for the third and fourth steps in Algorithm 1. To show the choice of the error bounds ε_i, $i = 0,1,2,3$ are satisfactory, we present the numerical results of L_0, L_1 and BP for large values of N in tables 1 – 3. For each case of γ in tables 1 – 3, the first line corresponds to the truncation level N obtained by the Algorithm 1 and the second line corresponds to the large N for comparisons.

Table 1: $M/M/25/K$ retrial queue with $\lambda = 1.0$ and $\rho = 0.9$

	$K=25$				$K=50$			
γ	N	L_0	L_1	BP	N	L_0	L_1	BP
10.	79	4.613	22.50	0.4984	54	0.3246	26.74	0.0355
	150	4.621	22.50	0.4984	100	0.3290	26.74	0.0356
1.	80	5.006	22.50	0.4479	54	0.3316	26.73	0.0307
	150	5.014	22.50	0.4479	100	0.3359	26.74	0.0308
0.5	81	5.3956	22.50	0.4166	54	0.3383	26.73	0.0276
	150	5.4037	22.50	0.4167	100	0.3425	26.73	0.0276
0.1	90	8.102	22.50	0.3273	89	0.3852	26.69	0.0176
	150	8.110	22.50	0.3274	100	0.3854	26.69	0.0176
0.05	120	11.24	22.50	0.2930	117	0.4309	26.64	0.0133
	150	11.24	22.50	0.2935	150	0.4309	26.64	0.0133

Table 2: $M/M/25/K$ retrial queue with $\lambda = 1.0$ and $\rho = 0.95$

	$K=25$				$K=50$			
γ	N	L_0	L_1	BP	N	L_0	L_1	BP
10.0	158	13.89	23.75	0.7199	133	3.793	33.74	0.1986
	300	13.94	23.75	0.7201	300	3.842	33.75	0.1991
1.0	160	14.75	23.75	0.6726	133	3.817	33.71	0.1805
	300	14.81	23.75	0.6728	300	3.866	33.72	0.1809
0.5	163	15.64	23.75	0.6410	133	3.841	33.69	0.1672
	300	15.70	23.75	0.6412	300	3.890	33.70	0.1676
0.1	179	22.02	23.75	0.5418	133	3.997	33.54	0.1174
	300	22.09	23.75	0.5420	300	4.045	33.54	0.1178
0.05	197	29.51	23.75	0.5005	134	4.164	33.38	0.0912
	300	29.60	23.75	0.5008	300	4.210	33.38	0.0915

Table 3: $M/M/25/K$ retrial queue with $\lambda = 1.0$ and $\rho = 0.975$

	$K=25$				$K=50$			
γ	N	L_0	L_1	BP	N	L_0	L_1	BP
10.0	292	33.27	24.38	0.8515	267	17.41	40.04	0.4502
	450	33.64	24.38	0.8519	400	17.74	40.06	0.4513
1.0	296	34.97	24.38	0.8164	267	17.45	40.00	0.4237
	450	35.34	24.38	0.8169	400	17.78	40.03	0.4249
0.5	301	36.76	24.38	0.7918	267	17.48	39.97	0.4027
	450	37.13	24.37	0.7923	400	17.81	39.99	0.4039
0.1	332	49.97	24.38	0.7096	267	17.70	39.74	0.3126
	450	50.39	24.38	0.7102	400	18.04	39.76	0.3139
0.05	365	65.72	24.37	0.6739	268	17.96	39.51	0.2566
	450	66.18	24.37	0.6746	400	18.28	39.53	0.2578

5 Concluding remarks

For the computation of the stationary distribution of LDQBD process, one of the main problem is to calculate the rate matrices R_k. Especially, when the truncation level N is large, the number of the boundary states are large and many inverse matrices $(-A_1^{(k)})^{-1}$ should be calculated for R_k. The algorithm presented in this paper follows the general method presented in Bright and Taylor [4], but main contribution of this paper is to present a simple algorithm for $(-A_1^{(k)})^{-1}$ when $A_1^{(k)}$ is the tridiagonal form. Thus the algorithm of this paper can be applied to the LDQBD process with generator whose diagonal block is tridiagonal form. Stepanov [13] considered a very general Markovian retrial queueing system with impatient customers and feedback, that is, an arriving customers from outside of the system and the orbit can leave the system without service and the customers who completes its service can rejoin the queue. Stepanov's model also falls into the LDQBD process with diagonal block that is tridiagonal matrix. Algorithm A for computing the inverse of the transient generator of tridiagonal form can be applied to the case where $A_1^{(k)}$ is the block tridiagonal form, for example $MAP/M/c/K$ retrial queue. In this case, using the block matrix version of algorithm A, it necessary to $K+1$ inverses of the matrix of size m instead of the matrix inverse of size $m(K+1)$ for the inverse of the diagonal block, where m is the size of the underlying Markov chain describing MAP (Markovian arrival process).

Appendix A Fundamental matrix of a generator of transient birth-and-death process with finite state space

Consider an $n \times n$ matrix T is of the form

$$T = \begin{pmatrix} b_1 & a_1 & & & & \\ c_2 & b_2 & a_2 & & & \\ & c_3 & b_3 & a_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{n-1} & b_{n-1} & a_{n-1} \\ & & & & c_n & b_n \end{pmatrix}.$$

with $b_i < 0$ for $1 \le i \le n$, and $a_i > 0$, $c_j > 0$ for $1 \le i \le n-1$, $2 \le j \le n$. We assume that $T\mathbf{1} \lneqq \mathbf{0}$ and hence T^{-1} exists, and denote by

$$(-T)^{-1} = X \equiv \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix}$$

the fundamental matrix of T.

Define w_k and v_k, $k = 1, 2, \cdots, n$ by the solution of the following equations

$$a_{k-1} + w_k b_k + w_k w_{k+1} c_{k+1} = -\delta_{1,k},\ 1 \le k \le n, \tag{A.1}$$
$$c_k + b_k v_k + a_k v_{k+1} v_k = -\delta_{1,k},\ 1 \le k \le n, \tag{A.2}$$

where $\delta_{i,j} = 1$ for $i = j$, $\delta_{i,j} = 0$ for $i \neq j$ and $a_0 = a_n = c_1 = c_{n+1} = 0$.

Lemma A.1. *The w_k and v_k, $1 \le k \le n$ are positive and given as follows*

$$w_n = a_{n-1}(-b_n)^{-1}, \tag{A.3}$$
$$w_k = a_{k-1}[-(b_k + w_{k+1} c_{k+1})]^{-1},\ k = n-1, n-2, \cdots, 2, \tag{A.4}$$
$$w_1 = [-(b_1 + w_2 c_2)]^{-1} \tag{A.5}$$

and

$$v_n = (-b_n)^{-1} c_n, \tag{A.6}$$
$$v_k = [-(b_k + a_k v_{k+1})]^{-1} c_k,\ k = n-1, n-2, \cdots, 2, \tag{A.7}$$
$$v_1 = [-(b_1 + a_1 v_2)]^{-1}. \tag{A.8}$$

Proof. The recursive formulae $(A.3) - (A.5)$ and $(A.6) - (A.8)$ are immediate from $(A.1)$ and $(A.2)$, respectively. It remains to show that $w_k > 0$ and $v_k > 0$, $1 \le k \le n$. It is clear that $w_n > 0$. We have from $T\mathbf{1} \lneqq \mathbf{0}$ and $c_{n-1} > 0$ that

$$-(b_{n-1} + w_n c_n) = -b_{n-1} + \frac{c_n}{b_n} a_{n-1} \ge -b_{n-1} - a_{n-1} \ge c_{n-1} > 0 \tag{A.9}$$

and hence

$$0 < -\frac{c_{n-1}}{b_{n-1} + w_n c_n} \le 1. \tag{A.10}$$

Similarly, it follows from $(A.9)$ and $(A.10)$ that

$$-(b_{n-2} + w_{n-1} c_{n-1}) = -b_{n-2} + \frac{c_{n-1} a_{n-2}}{b_{n-1} + w_{n-1} c_{n-1}} \ge -b_{n-2} - a_{n-2} \ge c_{n-2} > 0$$

and hence

$$0 < -\frac{c_{n-2}}{b_{n-2} + w_{n-1} c_{n-1}} \le 1.$$

Repeating this procedure, we have that $-(b_k + w_{k+1} c_{k+1}) > 0$ and $w_k > 0$, $1 \le k \le n-1$. Similarly, we can show $-(b_k + a_k v_{k+1}) > 0$ and hence $v_k > 0$, $1 \le k \le n-1$. □

Proposition A.2. *The (i, j)-component x_{ij}, $1 \le j \le n$ of $X = (-T)^{-1}$ are given as follows:*

(1) The first row and column :

$$x_{1j} = w_1 w_2 \cdots w_j,\ j = 1, 2, \cdots, n, \tag{A.11}$$
$$x_{j1} = v_j v_{j-1} \cdots v_1,\ i = 1, 2, \cdots, n, \tag{A.12}$$

(2) The j-th $(2 \le j \le n-1)$ *row and column :*

$$\begin{aligned}
x_{jj} &= (1+x_{j,j-1}a_{i-1})[-(b_i+w_{i+1}c_{i+1})]^{-1}, && (A.13)\\
x_{jm} &= x_{jj}w_{j+1}w_{j+2}\cdots w_m,\ m=j+1,j+2,\cdots,n, && (A.14)\\
x_{mj} &= v_m v_{m-1}\cdots v_{j+1}x_{jj},\ m=j+1,j+2,\cdots,n, && (A.15)
\end{aligned}$$

(3) The (n,n)*-component :*

$$x_{nn} = (1+x_{n,n-1}a_{n-1})[-b_n]^{-1}. \tag{A.16}$$

Proof. We have from $XT=-I$ and $TX=-I$ that

$$\begin{aligned}
x_{i,j-1}a_{j-1}+x_{ij}b_j+x_{i,j+1}c_{j+1} &= -\delta_{ij},\ 1\le i,j\le n, && (A.17)\\
c_i x_{i-1,j}+b_i x_{ij}+a_i x_{i+1,j} &= -\delta_{ij},\ 1\le i,j\le n,. && (A.18)
\end{aligned}$$

It can be seen from the construction of w_k and v_k that x_{ij}, $1\le i,j\le n$ given in $(A.11)$ – $(A.16)$ satisfy $(A.17)$ and $(A.18)$. □

Summarizing the results above, we present an algorithm for computing $(-T)^{-1}$.

Algorithm A. Computation of $(-T)^{-1}=(x_{ij})$

1. Compute w_k and v_k, $1\le k\le n$ using Lemma A.1.
2. Computation of rows and columns.
   For $i=1,2,\cdots,n$, do
   - $x_{jj}=\begin{cases} w_1, & j=1,\\ (1+x_{j,j-1}a_{j-1})[-(b_j+w_{j+1}c_{j+1})]^{-1}, & j\ge 2.\end{cases}$
   - $x_{jk}=x_{j,k-1}w_k$, $k=j+1,\cdots,n$: *j-th row*
   - $x_{kj}=v_k x_{k-1,j}$, $k=j+1,\cdots,n$: *j-th column*

Remark A.1. Carefully looking the proof of Proposition A.2, we can see that x_{jj} in $(A.13)$ may also be written by

$$x_{jj}=[-(b_j+a_j v_{j+1})]^{-1}(1+c_i x_{j-1,j}),\ \ j=2,3,\cdots,n. \tag{A.19}$$

Remark A.2. Replacing a_i, b_i and c_i by A_i, B_i and C_i, respectively, we can show that Proposition A.2 holds for the transient QBD generator T^* that is irreducible and is of the form

$$T^*=\begin{pmatrix}
B_1 & A_1 & & & & \\
C_2 & B_2 & A_2 & & & \\
 & C_3 & B_3 & A_3 & & \\
 & & \ddots & \ddots & \ddots & \\
 & & & C_{n-1} & B_{n-1} & A_{n-1}\\
 & & & & C_n & B_n
\end{pmatrix},$$

where B_k is the square matrix of order n_k, $1\le k\le n$, and A_k and C_k are the matrices with appropriate order.

References

[1] V. V. Anisimov and J. R. Artalejo. Analysis of Markov multiserver retrial queues with negative arrivals. *Queueing Systems*, 39, 157–182, 2001.

[2] V. V. Anisimov and J. R. Artalejo. Approximation of multiserver retrial queues by means of generalized truncation models. *Top*, 10, 51–66, 2002.

[3] J. R. Artalejo. A classified biliography of research on retrial queues: Progress in 1990–1999. *Top*, 7(2), 187–211, 1999.

[4] L. Bright and P. G. Taylor. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models*, 11, 497–525, 1995.

[5] G. I. Falin and J. G. C. Templeton. *Retrial Queues*, Chapman and Hall, London, 1997.

[6] Q. M. He, H. Li and Y. Q. Zhao. Ergodicity of the $BMAP/PH/s/s+K$ retrial queue with PH-retrial times. *Queueing Systems*, 35, 323–347, 2000.

[7] V. Naoumov. Matrix-multiplicative approach to quasi-birth-and-death processes analysis. *Matrix-Analytic Methods in Stochastic Models*, ed. by S. R. Chakravarthy and A. S. Alfa, Marcel Dekker, Inc., 87–106, 1997.

[8] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models — An Algorithmic Approach*, Johns Hopkins University Press, Baltimore, MD, 1981.

[9] M. F. Neuts and B. M. Rao. Numerical investigation of a multiserver retrial model. *Queueing Systems*, 7, 169–190, 1990.

[10] Y. W. Shin. Multi-server retrial queue with negative customers and disasters. Submitted.

[11] Y. W. Shin. Monotonicity properties in various retrial queues and their applications. To be appeared in *Queueing Systems*.

[12] Y. W. Shin and C. E. M. Pearce. An algorithmic approach to the Markov chain with transition probaility matrix of upper block-Hessenberg form. *Korean Journal of Computational and Applied Mathematics*, 5(2), 403–426, 1998.

[13] S. N. Stepanov. Generalized model with repeated calls in case of estreme load. *Queueing Systems*, 27, 131–151, 1997.