J. Appl. Math. & Informatics Vol. **35***(2017), No. 5 - 6, pp. 623 - 649*
https://doi.org/10.14317/jami.2017.623

DIMENSION REDUCTION FOR APPROXIMATION OF ADVANCED RETRIAL QUEUES : TUTORIAL AND REVIEW[†]

YANG WOO SHIN

Abstract. Retrial queues have been widely used to model the many practical situations arising from telephone systems, telecommunication networks and call centers. An approximation method for a simple Markovian retrial queue by reducing the two dimensional problem to one dimensional problem was presented by Fredericks and Reisner in 1979. The method seems to be a promising approach to approximate the retrial queues with complex structure, but the method has not been attracted a lot of attention for about thirty years. In this paper, we exposit the method in detail and show the usefulness of the method by presenting the recent results for approximating the retrial queues with complex structure such as multi-server retrial queues with phase type distribution of retrial time, impatient customers with general persistent function and/or multiclass customers, etc.

AMS Mathematics Subject Classification : 60K25, 68M20.
Key words and phrases : Retrial queue, multiclass of customers, impatient customers, phase type distribution of retrial time, dimension reduction method, approximation.

1. Introduction

Retrial queue is a queueing system with returning customers and it consists of service facility and a virtual buffer called orbit as depicted in Figure 1. The service facility behaves as a queueing system with finite buffer. Customers arrive from outside to the service facility and demand independent and identically distributed service. If there are available servers or available positions in waiting space upon arrival, the customer joins the service facility. On the other hand, if an arriving customer finds that all servers are busy and all the waiting positions are occupied, the customer leaves the system forever or try to get service again after random amount of time. Those customers who will come back and try to

Received May 6, 2017. Revised August 20, 2017. Accepted August 22, 2017.
[†]This research was supported by Changwon National University in 2017−2018.

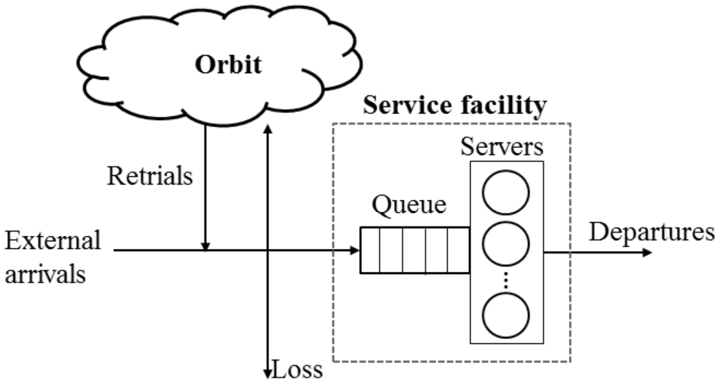

Figure 1. Schematic diagram of retrial queue

get service later are said to be in the virtual space called *orbit*. Unless otherwise mentioned, the capacity of orbit is assumed to be infinite. The customers in orbit retry independently to get service and are treated same as the primary customers that arrive from outside. That is, if a retrial customer finds free servers or waiting position available, the customer receives service immediately or joins the waiting space for service, otherwise, the customer leaves the system without service or joins orbit again. The time interval between retrials of a customer from orbit is called retrial time. The retrial times of a customer are assumed to be independent and identically distributed.

We denote the retrial queue by using the notation for usual queueing system. For example, $M/G/c/K$ retrial queue has a service facility with c identical servers in parallel and the capacity of service facility is K that includes the number of servers and waiting positions, the external arrival occurs according to a Poisson process and the service time of each server has general distribution. If there are no extra waiting positions in service facility, that is, $K = c$, then we omit the K and denote the system by $M/G/c$ retrial queue instead of $M/G/c/c$ retrial queue.

Retrial queues have been widely used to model the many practical situations arising from telephone systems, telecommunication networks and call centers. There are extensive literature about retrial queues. For details of application area, an overview and bibliographies of retrial queues, readers can refer to the survey papers [66, 17, 33, 10, 29], the survey of bibliographies [2, 3, 4, 22] and the books [19, 5].

Falin [17] divides the retrial queues into three large groups so called *single server systems*, *multi-server (fully available) systems* and *structurally complex systems* based on the nature of results obtained, methods of analysis and areas of applications. This classification is simplified by two groups *main models* and *advanced models* in [19]. The retrial queueing systems with exponential retrial time, no-loss system and the system with single class of customers are contained in the main model. On the other hand, the retrial queues with phase type (PH) distribution of retrial time, impatient customers, multi-class of customers, multi-server retrial queue with general distribution of interarrival time and/or the service time, and the server vacation are contained in the group of advanced models or structurally complex systems. In almost all the literature for retrial queues, analytic solutions have focused on single server system with Poisson arrival and computational or approximation methods have been developed for the main multi-server models. The monograph [19] presents details of analytical results for single server system and computational and approximation methods for multi-server system. The computational approaches are emphasized in the book [5] where various computational methods for single-server system and main $M/M/c$ retrial queues and matrix analytic method for the system with the structure of quasi-birth-and-death (QBD) process or the Markov chain of $M/G/1$ type are presented.

There are a few qualitative approach for the multi-server retrial queues in the third class, e.g. a stability condition for $BMAP/PH/s/s+K$ retrial queue with PH retrial time [25], a stability condition for $MAP/PH/s/K$ retrial queue with PH retrial time and server vacation [49] and $GI/G/c$ retrial queue with exponential retrial time [39] and the monotonicity for $M/G/1$ retrial queue with general retrial time [36] and for $A^X/B/c/K$ retrial queue with exponential retrial times [44]. The retrial queues with structurally complex structure have been known to be difficult for mathematical analysis because the joint queue length process is a random walk on the multidimensional integer lattice even it is modelled by a Markov chain. So, the explicit or computational approaches for these models are very limited. The truncation method or matrix analytic method in [19, 5] do not seem to be useful to obtain the performance measures about the multi-server retrial queue in advanced models.

Fredericks and Reisner [20] present an approximation method for multi-server retrial queue. The method reduces the two dimensional problem to one dimensional problem and makes the computation of stationary distribution become simpler and more efficient. The method has been also introduced as a method of reducing dimension [66], an approximation with the help of a loss model [19, Section 2.8.1] and Fredericks and Reisner approximation [5, Section 3.4.4]. However, the method has not been attracted a lot of attention until Shin and Moon [52] used it for an approximation of $M/M/c$ retrial queue with phase type distribution of retrial time. Here, we denote the Fredericks and Reisner's approach by *dimension reduction method.*

Recently, Shin and Moon present approximation results of the performance characteristics indicated in [17] for the complex systems by using the dimension reduction method in a series of papers [52, 53, 56, 58, 59, 60]. The objective this paper is to exposit the dimension reduction method and show the usefulness of the method by presenting the recent results of the authors.

The paper is organized as follows. In Section 2, the dimension reduction method is described with numerical results by approximating the simple $M/M/c$ retrial queue. Approximation of $PH/PH/c$ retrial queue with PH retrial times and its application to $GI/G/c$ retrial queue with general retrial time are presented in Section 3. Approximation of $M/M/c/K$ retrial queue with impatient customers and $M/M/c$ with multi-class of customers are presented in Sections 4 and 5, respectively. Approximations of $M/M/c$ retrial queue with server vacations in which the retrial times and vacation times are of PH distributions is proposed in Section 6. Concluding remarks are given in Section 7.

2. Dimension reduction method

2.1. Model and preliminary results. Consider the main $M/M/c$ retrial queue with arrival rate λ, the service rate μ of each server and retrial rate γ.

Let $X_0(t)$ be the number of customers at service facility and $X_1(t)$ be the number of customers in orbit at time t. Then $\boldsymbol{X} = \{(X_0(t), X_1(t)), t \geq 0\}$ is a continuous time Markov chain with state space $\mathcal{S} = \{0, 1, \cdots, c\} \times \mathbb{Z}_+$, where $\mathbb{Z}_+ = \{0, 1, 2, \cdots\}$. For stability of the system we assume that $\rho = \frac{\lambda}{c\mu} < 1$ [19]. Let (X_0, X_1) be the stationary version of $\boldsymbol{X}$ and $P(j, n) = P(X_0 = j, X_1 = n)$ for $(j, n) \in \mathcal{S}$ and $P(j, n) = 0$ otherwise. Following the standard procedure, the balance equations for the Markov chain $\boldsymbol{X}$ are easily obtained as follows: for $(j, n) \in \mathcal{S}$ with $0 \leq k < c$,

$$(\lambda + \mu_j + n\gamma)\, P(j, n) = \lambda P(j-1, n) + \mu_{j+1} P(j+1, n) + (n+1)\gamma P(j-1, n+1), \tag{1}$$

where $\mu_j = j\mu$ and for $j = c$,

$$(\lambda + c\mu)\, P(c, n) = \lambda P(c-1, n) + \lambda P(c, n-1) + (n+1)\gamma P(c-1, n+1). \tag{2}$$

Let $\pi_j = P(X_0 = j)$, $j = 0, 1, \cdots, c$. Summing over n in (1) yields that

$$(\lambda_j + \mu_j)\pi_j = \lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1},\ 0 \leq j < c, \tag{3}$$

where

$$\lambda_j = \lambda + \gamma L_j,\ j = 0, 1, \cdots, c-1 \tag{4}$$

and $L_j = \mathbb{E}[X_1 | X_0 = j]$. We can see from (3) that

$$\mu_{i+1}\pi_{i+1} = \lambda_i \pi_i,\ 0 \leq i < c \tag{5}$$

and hence

$$\pi_j = \pi_0 \prod_{i=1}^{j} \left(\frac{\lambda_{i-1}}{\mu_i} \right),\ j = 1, 2, \cdots, c$$

with

$$\pi_0 = \left[1 + \sum_{j=1}^{c} \prod_{i=1}^{j} \left(\frac{\lambda_{i-1}}{\mu_i} \right) \right]^{-1}.$$

Furthermore, summing over i in both sides of (5), we have that

$$\mu\mathbb{E}[X_0] = \sum_{i=0}^{c-1} \lambda_i \pi_i = \lambda(1-\pi_c) + \gamma(L - L_c\pi_c) = \lambda$$

and

$$\mathbb{E}[X_0] = \frac{\lambda}{\mu} \tag{6}$$

which is consistent with Little's law.

Note that π_j is expressed in terms of unknown λ_j. We need some preliminaries to determine unknowns. Let $P_n = P(X_1 = n)$, $n = 0, 1, 2, \cdots$. Summing over j in (1) and (2) yields that

$$\begin{aligned}(n+1)\gamma P_{n+1} - n\gamma P_n &= \lambda(P(c,n) - P(c,n-1)) \\ &+ \gamma((n+1)P(c,n+1) - nP(c,n))\end{aligned}$$

and hence

$$\gamma n P_n = \lambda P(c, n-1) + \gamma n P(c,n)$$

Thus the mean number $L = \mathbb{E}[X_1]$ of customers in orbit is given by

$$L = \lambda_c \pi_c m_r, \tag{7}$$

where $m_r = \frac{1}{\gamma}$ is the mean retrial time and $\lambda_c = \lambda + \gamma L_c$.

Let R_j be the proportion of returning customers from orbit who find the service facility in state j, that is, (e.g. see [64, page 370], [5, page 78], [19, page 169])

$$R_j = \frac{\gamma\mathbb{E}[X_1 1_{\{X_0=j\}}]}{\gamma\mathbb{E}[X_1]} = \frac{L_j\pi_j}{L},\ j = 0, 1, \cdots, c. \tag{8}$$

It follows from (7) and (8) that $\gamma L_j = \frac{\lambda_c\pi_c}{\pi_j} R_j$ and hence

$$\lambda_c = \frac{\lambda}{1 - R_c}. \tag{9}$$

Combining (7) and (9), we have that

$$L = m_r \frac{\lambda\pi_c}{1 - R_c} \tag{10}$$

Summarizing the results above with (4), we have that

$$\lambda_j = \lambda + \lambda \left(\prod_{i=j+1}^{c} \frac{\lambda_{i-1}}{\mu_i} \right) \frac{R_j}{1 - R_c},\ 0 \le j < c. \tag{11}$$

2.2. Approximation. Note that the equation (3) is the same as the balance equation of the birth-and-death process with birth rates λ_j and death rates μ_j. The λ_j represents the sum of the arrival rate λ from outside and the arrival rate γL_j from orbit to the service facility under the condition $X_0 = j$. It can be seen that $\{\lambda_j\}_{j=0}^{c-1}$ in (4) have already reflected the external arrival and retrials and $\{\mu_j\}_{j=1}^{c}$ are independent of retrials. Based on this observation, we adopt the following approximation assumption about the behavior of service facility for an approximation of λ_j.

Assumption M. *The service facility behaves like a birth and death process with birth rates λ_i, $0 \le i \le c$ in* (4) *and death rates μ_j, $0 \le j \le c$ and is independent of the retrials.*

Let Q be the infinitesimal generator of the birth and death process $\{\xi(t), t \ge 0\}$ with birth rates λ_i, $0 \le i \le c-1$ and death rates μ_j, $1 \le j \le c$ on the state space $\{0, 1, \cdots, c\}$ and $q_{ij}(t) = P(\xi(t) = j \,|\, \xi(0) = i)$. Assume that a customer is blocked to enter the service facility and joins orbit at time $t = 0$. This customer returns after exponential retrial time with rate γ and R_j is aproximated by the probability that the returning customer finds the service facility of state j under the Assumption M, that is,

$$R_j \approx \int_0^\infty q_{cj}(t)\gamma e^{-\gamma t}\,dt = \gamma[(\gamma I - Q)^{-1}]_{cj}, \; j = 0, 1, 2, \cdots, c. \tag{12}$$

where $[A]_{ij}$ is the (i,j)-component of the matrix A and I is the identity matrix.

The following iteration algorithm is for computing $\boldsymbol{\lambda} = (\lambda_0, \cdots, \lambda_c)$. We write Q as $Q(\boldsymbol{\lambda})$ to highlight the dependence of $\boldsymbol{\lambda}$.

Algorithm M

```
Step 0. [Initial step] Set $\lambda_j^{(0)} = \lambda$, $j = 0, 1, \cdots, c-1$.
Step 1. [Repeating step]. Repeat the following for $n = 1, 2, \cdots$
  1. Let $Q^{(n-1)} = Q(\boldsymbol{\lambda}^{(n-1)})$ and compute $R_j^{(n-1)}$ using (12)
  2. Update $\boldsymbol{\lambda}^{(n)}$ using (11) until
         $TOL = ||\boldsymbol{\lambda}^{(n)} - \boldsymbol{\lambda}^{(n-1)}|| < \epsilon$
for a given tolerance $\epsilon > 0$.
```

Although the convergence of the iteration scheme is not proved analytically, extensive numerical experiments show the convergence of the sequence $\{\boldsymbol{\lambda}^{(n)}\}_{n=0}^{\infty}$.

2.3. Performance measures and numerical results. As indicated in Falin [17] (see also [19]), the most important characteristics of the quality of service of customers in retrial queueing systems are *stationary blocking probability* $P_B = P(X_0 = c)$, *mean number* $L = \mathbb{E}[X_1]$ *of customers in orbit* and *the mean number* $\mathbb{E}[X_0]$ *of busy servers* from the practical point of view.

TABLE 1. P_B and σ_0 in $M/M/5$ retrial queue

			P_B			σ_0	
ρ	m_r	Exact	Appr.	Err(%)	Exact	Appr.	Err(%)
0.4	0.1	0.0554	0.0545	1.67	1.364	1.363	0.13
	1.0	0.0469	0.0460	1.90	1.341	1.338	0.25
	5.0	0.0432	0.0429	0.72	1.326	1.324	0.13
	10.0	0.0425	0.0424	0.40	1.322	1.321	0.07
	20.0	0.0422	0.0421	0.21	1.321	1.320	0.04
0.8	0.1	0.5215	0.4912	5.80	1.291	1.243	3.67
	1.0	0.4553	0.4299	5.59	1.170	1.113	4.86
	5.0	0.4263	0.4181	1.91	1.105	1.086	1.77
	10.0	0.4210	0.4166	1.04	1.092	1.082	0.97
	20.0	0.4181	0.4158	0.54	1.086	1.080	0.51

TABLE 2. L in $M/M/5$ retrial queue with $\rho = 0.6$

m_r	0.0	0.1	1.0	5.0	10.0	20.0
Exact	0.3542	0.4287	0.9820	3.2953	6.1727	11.9247
Appr.	0.1417*	0.2429	0.8078	3.1192	5.9962	11.7481
Error	0.2125	0.1858	0.1742	0.1761	0.1765	0.1766
Err(%)		43.3	17.7	5.3	2.9	1.5

* Approximation results are for $m_1 = 10^{-4}$

The blocking probability P_B and $\mathbb{E}[X_0]$ are immediately obtained from $\{\pi_j\}$ and L is given by (10). In Tables 1 - 3, numerical results are presented for the performance of approximations for $M/M/5$ retrial queue with $\mu = 1.0$ and the arrival rate $\lambda = 5\rho$. Numerical experiments provide that approximations for $\mathbb{E}[X_0]$ is the same as the exact results (6). In Table 1, approximation results for P_B and the standard deviation $\sigma_0 = \sqrt{\mathrm{Var}[X_0]}$ of X_0 are compared with the exact one, where the exact results are in fact calculated by the generalized truncation method, see [5, 19, 41]. The relative error is given by Err(%) $= \frac{(\text{Exact}-\text{Appr.})}{\text{Exact}} \times 100$. Table 1 shows that the approximation works well especially for large mean retrial time $m_r = \frac{1}{\gamma}$ and the approximation underestimates the exact one.

In Table 2, the exact results for $m_r = 0$ are for the ordinary $M/M/5$ queue and the corresponding approximation results are for $m_r = 10^{-4}$. We can see from Table 2 that the approximation of L does not provide satisfactory accuracy for small value of m_r. For example, in case of $m_r = 0.1$, the relative error Err(%) is 43.3%. However, the differences $\text{Error}(m_r) = L(m_r) - L_{\text{Appr}}(m_r)$ between exact one $L(m_r)$ and approximation result $L_{\text{Appr}}(m_r)$ for L with m_r slowly change as

TABLE 3. $\hat{L}$ and L in $M/M/5$ retrial queue

	$\rho = 0.4$			$\rho = 0.8$		
m_r	Exact	Appr.	Err(%)	Exact	Appr.	Err(%)
0.1	0.0509	0.0532	4.72	2.567	2.702	5.26
1.0	0.1345	0.1399	4.05	5.261	5.317	1.05
5.0	0.4851	0.4920	1.41	16.672	16.663	0.06
10.0	0.9213	0.9285	0.78	30.871	30.849	0.07
20.0	1.7933	1.8007	0.41	59.253	59.224	0.05

m_r increases. We propose a modified formula for $L(m_r)$ by

$$\hat{L}(m_r) = L_{\mathrm{Appr}}(m_r) + (L_{M/M/c} - L_{\mathrm{Appr}}(m_r^*)), \tag{13}$$

where $L_{M/M/c}$ is the mean number of customers in queue for the ordinary $M/M/c$ queue and m_r^* is sufficiently small value, e.g. $m_r^* = 10^{-4}$. Approximation results $\hat{L}$ with (13) are compared with exact ones in Table 3.

2.4. Bibliographical notes. Greenberg and Wolff [23] present an approximation for the stationary distribution of the number of customers in service facility in the $M/M/c/K$ retrial queue under the assumption that retrials see time averages (RTA). The approximation using RTA assumption does not reflect the retrial rate and works well only for small value of retrial rate. The dimension reduction method reflects the retrial rate and it provides more accurate approximation results than RTA approximation as the retrial rate increases.

3. $PH/PH/c$ retrial queue with PH retrial time

In this section, we apply the dimension reduction method to the retrial queue with phase type (PH) distribution of retrial times. The results of this section are from [53, 58].

3.1. The model. Consider the $PH/PH/c$ retrial queue with PH retrial time. The interarrival time, service time and retrial time are of PH-distributions with representation $PH(\boldsymbol{\alpha},\boldsymbol{T})$ (interarrival time), $PH(\boldsymbol{\beta},\boldsymbol{S})$ (service time) and $PH(\boldsymbol{\theta},\boldsymbol{U})$ (retrial time), where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_l)$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)$ and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_\nu)$ are row vectors of size l, m and ν, respectively and $\boldsymbol{T} = (t_{ij})$, $\boldsymbol{S} = (s_{ij})$ and $\boldsymbol{U} = (u_{ij})$ are the square matrices of size l, m and ν, respectively. The mean interarrival time, mean service time and mean retrial time are given by $m_a = \boldsymbol{\alpha}(-\boldsymbol{T})^{-1}\mathbf{e}$, $m_s = \boldsymbol{\beta}(-\boldsymbol{S})^{-1}\mathbf{e}$ and $m_r = \boldsymbol{\theta}(-\boldsymbol{U})^{-1}\mathbf{e}$, respectively, where $\mathbf{e}$ is the column vector of appropriate size whose components are all 1. For details of the PH-distribution and PH-renewal process, see [40, Chapter 2]. The stability condition of the system is $\rho = \frac{m_s}{cm_a} < 1$ [25, 49]. Let $\boldsymbol{\lambda} = -\boldsymbol{T}\mathbf{e}$, $\boldsymbol{\mu} = -\boldsymbol{S}\mathbf{e}$ and $\boldsymbol{\gamma} = -\boldsymbol{U}\mathbf{e}$ and denote the jth component of $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ by λ_j, μ_j and γ_j, respectively. Let $\boldsymbol{t} = (t_1, \cdots, t_l)$ with $t_i = -t_{ii}$ and similarly denote by $\boldsymbol{s} = (s_1, \cdots, s_m)$ and $\boldsymbol{u} = (u_1, \cdots, u_\nu)$ with $s_i = -s_{ii}$ and $u_i = -u_{ii}$. For later use, define some notation for vectors $\boldsymbol{x} = (x_1, \cdots, x_n)$ and $\boldsymbol{y} = (y_1, \cdots, y_n)$ by $|\boldsymbol{x}| = \sum_{i=1}^n x_i$, $\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^n x_i y_i$. Let $\mathbf{e}_i$ be the vector of appropriate size whose ith component is 1 and others are all 0 and denote the (k, k')-component of the matrix M by $[M]_{k,k'}$.

3.2. Stationary equations. Let $J(t)$ be the phase of arrival process, $X_i(t)$ the number of customers at service facility whose service phase is of i and $Y_j(t)$ be the number of customers in orbit whose retrial phase is of j at time t and $\boldsymbol{X}(t) = (X_1(t), \cdots, X_m(t)), \boldsymbol{Y}(t) = (Y_1(t), \cdots, Y_\nu(t))$. Then $\boldsymbol{\Psi} = \{(J(t), \boldsymbol{X}(t), \boldsymbol{Y}(t)), t \geq 0\}$ is a continuous time Markov chain on the state space $\mathcal{S} = \{1, \cdots, l\} \times \mathcal{K} \times \mathbb{Z}_+^\nu$, where $\mathcal{K} = \cup_{i=0}^c \mathcal{K}(i)$ and $\mathcal{K}(i) = \{(k_1, \cdots, k_m) \in \mathbb{Z}_+^m \ : \ \sum_{j=1}^m k_j = i\}$.

Let $(J, \boldsymbol{X}, \boldsymbol{Y})$ be the stationary version of $(J(t), \boldsymbol{X}(t), \boldsymbol{Y}(t))$. Following the usual arguments, one can see that the balance equations for $P(j, \boldsymbol{k}, \boldsymbol{n}) = P(J = j, \boldsymbol{X} = \boldsymbol{k}, \boldsymbol{Y} = \boldsymbol{n})$ from which the marginal distribution of $(J, \boldsymbol{X})$ are obtained as the following Propositions 3.1 and 3.2.

Proposition 3.1. *Let* $\pi(j, \boldsymbol{k}) = P(J = j, \boldsymbol{X} = \boldsymbol{k})$ *and*

$$a(j, \boldsymbol{k}) = \sum_{i=1}^{\nu} \gamma_i L_i(j, \boldsymbol{k}), \ 1 \leq j \leq l, \boldsymbol{k} \in \mathcal{K},$$

where

$$L_i(j, \boldsymbol{k}) = \mathbb{E}[Y_i | J = j, \boldsymbol{X} = \boldsymbol{k}].$$

Then $\boldsymbol{\pi} = (\pi(j, \boldsymbol{k}), j = 1, \cdots, l, \boldsymbol{k} \in \mathcal{K})$ *satisfies* $\boldsymbol{\pi} Q_{PH} = 0$, *where*

$$Q_{PH} = \begin{pmatrix} B_0 & A_0 & & & \\ C_1 & B_1 & A_1 & & \\ & \ddots & \ddots & \ddots & \\ & & C_{c-1} & B_{c-1} & A_{c-1} \\ & & & C_c & B_c \end{pmatrix}.$$

The matrix $B_k = (B_k(i,j))_{1 \leq i,j \leq l}$ *is a square matrix of size* $l\binom{m+k-1}{m-1}$, *where the block component* $B_k(i,j)$ *is a square matrix of size* $\binom{m+k-1}{m-1}$, $k = 0, 1, \cdots, c$ *whose components are as follows:*

$$B_k(i,j) = \begin{cases} t_{ij} I, & i \neq j, \ 0 \leq k \leq c-1 \\ (t_{ij} + \lambda_i \alpha_j) I, & i \neq j, \ k = c, \end{cases}$$

$$[B_k(i,i)]_{\boldsymbol{k}\boldsymbol{k}} = \begin{cases} -\left(t_i + a(i, \boldsymbol{k}) + \boldsymbol{k} \cdot \boldsymbol{s}\right), & 0 \leq k \leq c-1 \\ -\left(t_i - \lambda_i \alpha_i + \boldsymbol{k} \cdot \boldsymbol{s}\right), & k = c, \end{cases}$$

$$[B_k(i,i)]_{\boldsymbol{k}\boldsymbol{k}'} = \begin{cases} k_h s_{hj}, & \boldsymbol{k}' = \boldsymbol{k} + \mathbf{e}_j - \mathbf{e}_h, \ 0 \leq k \leq c, \\ 0, & \textit{otherwise}, \end{cases} \ \boldsymbol{k}' \neq \boldsymbol{k}.$$

The block matrix components of the matrices $A_k = (A_k(i,j))_{1 \leq i,j \leq l}$ *and* $C_k = (C_k(i,j))_{1 \leq i,j \leq l}$ *are as follows:*

$$[A_k(i,j)]_{\boldsymbol{k}\boldsymbol{k}'} = \begin{cases} (\lambda_i \alpha_j + a(i, \boldsymbol{k}) \delta_{ij}) \beta_h, & \boldsymbol{k}' = \boldsymbol{k} + \mathbf{e}_h, \ 0 \leq k \leq c-1 \\ 0, & \textit{otherwise}, \end{cases}$$

$$[C_k(i,j)]_{\boldsymbol{k}\boldsymbol{k}'} = \begin{cases} k_h\mu_h\delta_{ij}, & \boldsymbol{k}' = \boldsymbol{k} - \mathbf{e}_h,\ 1 \le k \le c, \\ 0, & otherwise, \end{cases}$$

where $\delta_{ij} = 1$ for $i = j$ and 0 otherwise. Note that the matrices of $A_k(i,j)$ and $C_k(i,j)$ are of size $\binom{m+k-1}{m-1} \times \binom{m+k}{m-1}$ and $\binom{m+k-1}{m-1} \times \binom{m+k-2}{m-1}$, respectively.

Proposition 3.2. *Let $\pi(j,\mathcal{K}(c)) = P(J = j, \boldsymbol{X} \in \mathcal{K}(c))$ and*

$$\Lambda = \sum_{j=1}^{l} \lambda_j \pi(j,\mathcal{K}(c)) + \sum_{i=1}^{\nu} \gamma_i \mathbb{E}[Y_i 1_{\{\boldsymbol{X}\in\mathcal{K}(c)\}}].$$

Then the vector $\boldsymbol{L} = (L_1, \cdots, L_\nu)$ with $L_i = \mathbb{E}[Y_i]$ is given by

$$\boldsymbol{L} = \Lambda\boldsymbol{\theta}(-\boldsymbol{U})^{-1},$$

$$L = \sum_{i=1}^{\nu} L_i = \Lambda m_r.$$

Note that $a(j,\boldsymbol{k})$ is the mean retrial rate from orbit given that the arrival process and service facility is in state $(j,\boldsymbol{k})$ and Λ is the total arrival rate to orbit. It follows from Proposition 3.1 that the service facility behaves like a $PH/PH/c/c$ loss system with extra arrivals from orbit with rate $a(j,\boldsymbol{k})$ that depends on the state in stationary state.

Remark 3.1. For the $M/M/c$ retrial queue with PH-retrial time, the generator Q_{PH} is the same as that of birth and death process with birth rates

$$\lambda_j = \lambda + \sum_{i=1}^{\nu} \gamma_i E[Y_i \,|\, X_0 = j],\ j = 0, 1, \cdots, c-1$$

and death rate $\mu_j = j\mu$, $j = 1, 2, \cdots, c$, where μ is the service rate and it can be easily seen that $L_0 = \sum_{i=1}^{m} \mathbb{E}[X_i] = \frac{m_s}{m_a}$, see [53].

Proposition 3.3. *Let $R(i,\boldsymbol{k})$ be the proportion of returning customers from orbit who find the arrival phase and service facility in state $(i,\boldsymbol{k})$, that is,*

$$R(j,\boldsymbol{k}) = \frac{\sum_{i=1}^{\nu} \gamma_i \mathbb{E}[Y_i 1_{\{J=j,\boldsymbol{X}=\boldsymbol{k}\}}]}{\sum_{i=1}^{\nu} \gamma_i L_i}.$$

Then

$$a(j,\boldsymbol{k}) = \Lambda R(j,\boldsymbol{k})/\pi(j,\boldsymbol{k}),\ \ \boldsymbol{k} \in \mathcal{K},\ 1 \le j \le l,$$

$$\Lambda = \frac{\lambda^* P_B^o}{1 - R_B},$$

where the proportion R_B that returning customers are blocked and the probability P_B^o that an arriving customer is blocked are as follows

$$R_B = \sum_{j=1}^{l} \sum_{\boldsymbol{k}\in\mathcal{K}(c)} R(j,\boldsymbol{k}),$$

$$P_B^o \quad = \quad \frac{1}{\lambda^*}\sum_{j=1}^{l}\lambda_j \pi(j, \mathcal{K}(c)),$$

where $\lambda^* = \frac{1}{m_a}$ *is the arrival rate.*

3.3. Approximations. Based on the formula of Q_{PH}, we adopt the following approximation assumption about the behavior of service facility.

Assumption PH. *The service facility behaves like a level dependent quasi-birth-and-death process with generator* Q_{PH} *and is independent of the retrials.*

In order to approximate $R(j, \boldsymbol{k})$, we classify the customers in orbit into two types. Let the customers who have experienced no retrial be of type 0 and the customers who have experienced one or more retrials be of type 1. Let $R_i(j, \boldsymbol{k})$, $i = 0, 1$ be the probability that a type i customer finds the system in $(j, \boldsymbol{k})$ upon retrial. The portion of type 0 customers in orbit is $\frac{\lambda^* P_B^o}{\Lambda} = 1 - R_B$. We approximate $R(j, \boldsymbol{k})$ by

$$R(j, \boldsymbol{k}) \approx (1 - R_B)R_0(j, \boldsymbol{k}) + R_B R_1(j, \boldsymbol{k}),\ 1 \leq j \leq l,\ \boldsymbol{k} \in \mathcal{K}. \qquad (14)$$

Assume that an arriving customer from outside is blocked at time $t = 0$ and the customer returns after $PH(\boldsymbol{\theta}, \boldsymbol{U})$ time. Under the Assumption PH, $R_0(j, \boldsymbol{k})$ is approximated by the probability that this customer finds the system in $(j, \boldsymbol{k})$ upon retrial as follows:

$$R_0(j, \boldsymbol{k}) \approx \frac{1}{P_B^o}\sum_{i=1}^{l}\sum_{\boldsymbol{k}' \in \mathcal{K}(c)} \frac{\lambda_i \pi(i, \boldsymbol{k}')}{\lambda^*}\left[\int_0^\infty \exp(Q_{PH}t)\boldsymbol{\theta} e^{\boldsymbol{U}t}\boldsymbol{\gamma} dt\right]_{(i,\boldsymbol{k}'),(j,\boldsymbol{k})}.$$

Similarly, $R_1(j, \boldsymbol{k})$ is approximated by

$$R_1(j, \boldsymbol{k}) \approx \frac{1}{P_B}\sum_{i=1}^{l}\sum_{\boldsymbol{k}' \in \mathcal{K}(c)} \pi(i, \boldsymbol{k}')\left[\int_0^\infty \exp(Q_{PH}t)\boldsymbol{\theta} e^{\boldsymbol{U}t}\boldsymbol{\gamma} dt\right]_{(i,\boldsymbol{k}'),(j,\boldsymbol{k})}.$$

Remark 3.2. If the arrival process is a Poisson process with rate λ, then it can be seen that

$$P_B^o = P(\boldsymbol{X} \in \mathcal{K}(c)) = P_B$$

and the proportion of returning customers from orbit who find the service facility in state $\boldsymbol{k}$ is approximated by

$$R(\boldsymbol{k}) \approx \frac{1}{P_B}\sum_{i=1}^{l}\sum_{\boldsymbol{k}' \in \mathcal{K}(c)} \pi(i, \boldsymbol{k}')\left[\int_0^\infty \exp(Q_{PH}t)\boldsymbol{\theta} e^{\boldsymbol{U}t}\boldsymbol{\gamma} dt\right]_{(i,\boldsymbol{k}'),(j,\boldsymbol{k})}.$$

The following algorithm is for computing $a(j, \boldsymbol{k})$ and $\pi(j, \boldsymbol{k})$. Let $\boldsymbol{a} = (a(j, \boldsymbol{k}), 1 \leq j \leq l, \boldsymbol{k} \in \mathcal{K})$ and we write Q_{PH} as $Q_{PH}(\boldsymbol{a})$ to highlight the dependence of $\boldsymbol{a}$.

Algorithm PH
Repeat the following steps starting with $\boldsymbol{a}^{(0)} = 0$ and $Q^{(0)} = Q_{PH}(\boldsymbol{a}^{(0)})$;

for $n = 0, 1, 2, \cdots$

1. Compute the stationary distribution $\boldsymbol{\pi}^{(n)}$ of $Q^{(n)}$
2. Compute $R^{(n)}(j, \boldsymbol{k})$ using (14) and then $\Lambda^{(n)}$
3. Update $Q^{(n+1)} = Q_{PH}(\boldsymbol{a}^{(n+1)})$ with $\boldsymbol{a}^{(n+1)} = \Lambda^{(n)} R^{(n)}(j, \boldsymbol{k}) / \pi^{(n)}(j, \boldsymbol{k})$

until

$$||\boldsymbol{a}^{(n+1)} - \boldsymbol{a}^{(n)}|| < \epsilon.$$

Remark 3.3. 1. Once $a(j, \boldsymbol{k})$ are determined, the stationary distribution $\boldsymbol{\pi}$ of Q_{PH} can be computed by the well known algorithm (e.g. see [45])

$$\pi(\mathcal{K}(n)) = \pi(\mathbf{0}) \mathcal{R}_1 \cdots \mathcal{R}_n, \ k = 1, 2, \cdots, c$$

where

$$\pi(\mathbf{0}) = \left(1 + \sum_{n=1}^{c} \mathcal{R}_1 \cdots \mathcal{R}_n \mathbf{e} \right)^{-1}$$

and $\boldsymbol{\pi}(\mathcal{K}(n)) = (\pi(j, \boldsymbol{k}), (j, \boldsymbol{k}) \in \{1, 2, \cdots, l\} \times \mathcal{K}(n))$. The matrices $\mathcal{R}_n$ of size $\binom{m+n-2}{m-1} \times \binom{m+n-1}{m-1}$ are calculated recursively as

$$\begin{aligned} \mathcal{R}_c &= (-B_c)^{-1} A_{c-1}, \\ \mathcal{R}_n &= A_{n-1} [-(B_n + \mathcal{R}_{n+1} C_{n+1})]^{-1}, \ n = c-1, c-2, \cdots, 1. \end{aligned}$$

2. The integration in $R_0(j, \boldsymbol{k})$ can be computed using the method in [53] as follows. Note that the Laplace-Stieltjes transform (LST) $\tilde{F}(\omega) = \boldsymbol{\theta}(\omega I - \boldsymbol{U})^{-1} \boldsymbol{\gamma}$ of the retrial time distribution $F(t)$ is a rational function and the probability density function $f(t)$ of $F(t)$ can be expressed by a linear combination of the function of the form $t^k e^{-\eta t}$ [62, Appendix E]. Thus $R(j, \boldsymbol{k})$ is the linear combination of

$$H(k, \eta) = \int_0^\infty \exp(Q_{PH} t) t^k e^{-\eta t} \, dt.$$

It can be easily seen that

$$H(k, \eta) = k! [(\eta I - Q_{PH})^{-1}]^{k+1}, \ k = 0, 1, \cdots.$$

One can use the algorithm in [45] for computing the inverse matrix $(\eta I - Q_{PH})^{-1}$.

3. Computing procedure can be interpreted as follows. Setting $\boldsymbol{a}^{(0)} = 0$ denotes that the retrial phenomena is ignored and $Q^{(0)}$ is the generator of ordinary $PH/PH/c/c$ loss system. Since $a^{(n)}(j, \boldsymbol{k}) = \sum_{i=1}^{\nu} \gamma_i L_i^{(n-1)}(j, \boldsymbol{k})$ denotes the arrival rate from the group of blocking customers in the system $Q^{(n-1)}$, the system $Q^{(n)}$ is the ordinary $PH/PH/c/c$ loss system with extra arrival rates $a^{(n)}(j, \boldsymbol{k})$. The convergence of Algorithm PH has not been proved analytically, but [58] reported the the convergence of the sequence $\{\boldsymbol{a}^{(n)}\}_{n=0}^{\infty}$ through extensive numerical experiments.

3.4. Performance measures. The performance measures such as blocking probability $P_B = P(\boldsymbol{X} \in \mathcal{K}(c))$, mean number of busy servers $L_0 = \sum_{i=1}^{m} \mathbb{E}[X_i]$ and the mean number of customers in orbit $L = \sum_{i=1}^{\nu} \mathbb{E}[Y_i]$ can be obtained from the approximation results of Λ and $\pi(j, \boldsymbol{k})$. By Little's law it can be easily seen that $L_0 = m_s/m_a$. In order to improve the accuracy of approximation for L, a modified formula [58]

$$\hat{L}(m_r) = L_{\mathrm{App}}(m_r) + (L_{PH/PH/c} - L_{\mathrm{Appr}}(m_r^*))$$

can be used, where $L(m_r)$ and $L_{\mathrm{App}}(m_r)$ are the exact and approximation of L as a function of m_r, and $L_{PH/PH/c}$ is the mean number of customers in queue for the ordinary $PH/PH/c$ queue and m_r^* is chosen to be small enough so that the variation of $L_{\mathrm{App}}(m_r)$ is negligible for $m_r \leq m_r^*$. Approximation formulae for the distribution of the number N_R of retrials made by a customer during its sojourn time in the system and LST of the sojourn time W during which a customer stays in the system until the customer leaves the system are given [58]. Here, we present the means of them as

$$\begin{aligned}\mathbb{E}[N_R] &= \frac{P_B^o}{1-R_B},\\ \mathbb{E}[W] &= m_s + \frac{P_B^o}{1-R_B} m_r = \frac{1}{m_a}(L_0 + L).\end{aligned}$$

Extensive numerical experiments in [53, 58] show that the approximation provides satisfactory accuracy when the squared coefficient of variation of retrial time $C_r^2 \leq 1$ or $C_r^2 > 1$ with $H = \frac{m_1 m_3}{3m_2^2/2} > 1$, where m_k, $k = 1, 2, 3$ is the kth moment of retrial time. In the case of $C_r^2 > 1$ and $H < 1$, the performance of approximation for P_B, σ_0 seems to be good, but the approximation $\hat{L}$ of L can be worse as ρ and C_r^2 increase.

Remark 3.4. 1. It is well known that the set of PH-distributions is dense (in the sense of weak convergence) in the set of all probability distributions on $(0, \infty)$ (e.g. see [6, page 84]). There are many moment matching methods for fitting the general distribution by the PH distributions e.g. [7, 26, 42, 62, 63].

One can use $PH/PH/c$ retrial queue for $GI/G/c$ retrial queue by approximating the distribution of interarrival time and service time with PH distributions based on the result. In order to choose an appropriate PH distribution, the sensitivity of the performance measures with respect to the moments of interarrival time, service time and retrial time is investigated in [54, 57]. Shin and Moon [58] choose $PH/PH/c$ retrial queue as an approximate model of $GI/G/c$ retrial queue by fitting the first three moments of interarrival time, service time and retrial time with PH distributions. Then, they approximate the $PH/PH/c$ retrial queue by using the dimension reduction method described in the previous section.

2. One should note the following comments in [57] when one uses the $PH/PH/c$ retrial queue for approximating $GI/G/c$ retrial queue. The matrix analytic

method requires long computation time and large size of memory when both of the number of phases in PH distribution and the number c of servers are large. So the method is limited to small values of c and to PH distribution of lower order. The method to approximate the multi-server queue by fitting the general distribution with PH distributions is not free from the restriction of matrix analytic method, e.g. [55].

3.5. Bibliographical notes. The literature about the retrial queues with non-exponential retrial time is very limited. The order relation for $GI/G/1$ retrial queue with PH-retrial time is considered in [37]. The stability condition for $BMAP/PH/c/K$ retrial queue with PH-retrial times are presented by [25]. Breuer et al. [9] indicate that the proof of the sufficient condition for $K = c$ is not correct. However, Shin [49] showed that the result of [25] is correct . Approximation methods for the system with non-exponential retrial times are developed for $M/G/1$ retrial queue with general retrial time [67], for $M/G/1$ retrial queue with mixture of Erlang retrial time [35] and for $M/PH/1$ retrial queue with PH-retrial time [15]. Algorithmic solution for $M/M/c$ retrial queue with PH_2 retrial time is developed in [46].

4. $M/M/c$ retrial queue with impatient customers

4.1. The model and preliminaries. Consider an $M/M/c/K$ retrial queue with arrival rate λ and the service rate μ of each server, where the customers are impatient. The impatience of customers is governed by the persistence function $\{H_k, k = 1, 2, \cdots\}$, where H_k is the probability that a customer will join orbit after kth fail to enter the service facility. For a technical reason, we assume that the number of retrials of a customer from orbit is limited by m, that is, $H_k = 0$ for $k \geq m + 1$. Since $H_k = 0$ for $k \geq m + 1$, it can be easily seen that the system is always stable. We assume that the retrial rate may depend on the number of failures to enter the service facility and let γ_k be the retrial rate of the customer that has experienced blocking k times. An approximation for this system is proposed in [56] and we describe the method and the results therein briefly in the following.

Denote the customer who has failed k times to enter the service facility and is still in orbit by the type k-customer, $1 \leq k \leq m$. Thus if a type k-customer is blocked again, then the customer becomes the customer of type $k + 1$ with probability H_{k+1} or leaves the system with probability $\bar{H}_{k+1} = 1 - H_{k+1}$, $1 \leq k < m$. The customers in service facility are denoted by type 0 ones. Let $X_k^{(m)}$ be the number of customers of type k, $k = 0, 1, \cdots, m$ in stationary state. We fix a finite number m and for simplicity, we write X_k instead of $X_k^{(m)}$ if it is not confused in the context and let $P_{kj} = P(X_k = j)$ be the distribution of X_k, $0 \leq k \leq m$.

Note that the state space of the random vector $\boldsymbol{X} = (X_0, \cdots, X_m)$ is $\mathcal{S} = \{0, 1, \cdots, K\} \times \mathbb{Z}_+^m$. Following the usual arguments, one can easily derive the the balance equations for $P(\boldsymbol{n}) = P(\boldsymbol{X} = \boldsymbol{n})$. It can be easily seen from the balance equations that the marginal distribution P_{0j} of X_0 satisfies the following equations

$$(\lambda_j + \mu_j)P_{0j} = \lambda_{j-1}P_{0,j-1} + \mu_{j+1}P_{0,j+1},\ 0 \leq j < K, \tag{15}$$

where $\mu_j = \min(j, c)\mu$ and

$$\lambda_j = \lambda + \sum_{k=1}^{m} \gamma_k L_{kj},\ j = 0, 1, \cdots, K-1 \tag{16}$$

with

$$L_{kj} = \mathbb{E}[X_k | X_0 = j],\ 0 \leq k \leq m,\ 0 \leq j \leq K.$$

Thus P_{0j}, $j = 0, 1, \cdots, K$ are the stationary distribution of birth and death process with arrival rates λ_j and service rates μ_j and the following proposition is immediately obtained.

Proposition 4.1. *The marginal distribution $P_{0j} = P(X_0 = j)$ is given by*

$$P_{0j} = P_{00} \prod_{i=1}^{j} \left(\frac{\lambda_{i-1}}{\mu_i}\right),\ j = 1, 2, \cdots, c \tag{17}$$

with $\sum_{j=0}^{c} P_{0j} = 1$.

We can also obtain from the balance equations for $P(\boldsymbol{n})$ that the mean $L_k = \mathbb{E}[X_k]$ is given by

$$L_k = \begin{cases} \frac{1}{\gamma_1}\lambda P_{0K}H_1, & k = 1, \\ \frac{1}{\gamma_k}\gamma_{k-1}L_{k-1,K}P_{0K}H_k, & k = 2, 3, \cdots, m. \end{cases} \tag{18}$$

Let R_{kj} be the proportion of returning customers of type k who find the service facility of state j, that is,

$$R_{kj} = \frac{\gamma_k \mathbb{E}[X_k 1_{\{X_0=j\}}]}{\gamma_k \mathbb{E}[X_k]} = \frac{L_{kj}P_{0j}}{L_k},\ k = 1, 2, \cdots, m. \tag{19}$$

It follows from (18), (19) and (17) that

$$L_{kj} = \begin{cases} \frac{\lambda}{\gamma_k} \prod_{i=1}^{k} H_i R_{iK}, & j = K, \\ \frac{\lambda}{\gamma_k} \left(\prod_{i=j+1}^{K} \frac{\lambda_{i-1}}{\mu_i}\right) \left(\prod_{i=1}^{k-1} H_i R_{iK}\right) H_k R_{kj}, & 1 \leq j \leq K-1 \end{cases} \tag{20}$$

Combining the results above, we have that

Proposition 4.2.

$$L_k = \frac{\lambda P_{0K}}{\gamma_k} \left(\prod_{i=1}^{k-1} H_i R_{iK}\right) H_k,\ k = 1, 2, \cdots, m, \tag{21}$$

$$\lambda_j = \lambda + \lambda \left(\prod_{i=j+1}^{K} \frac{\lambda_{i-1}}{\mu_i}\right) \sum_{k=1}^{m} \left(\prod_{i=1}^{k-1} H_i R_{iK}\right) H_k R_{kj},\ 0 \leq j < K. \tag{22}$$

4.2. Approximations. Based on the formula (15), we adopt the following approximation assumption about the behavior of service facility.

Assumption IC. The service facility behaves like a birth-and-death process with birth rates $\{\lambda_j\}_{j=0}^{K-1}$ in (16) and death rates $\{\mu_j\}_{j=1}^{K}$ and is independent of the retrials.

Assume that a type $(k-1)$-customer is blocked at time $t=0$ and it becomes a type k-customer. This customer returns after an exponential time with parameter γ_k. Based on **Assumption IC**, we approximate R_{kj} with the probability that the returning customer finds the service facility of state j, that is,

$$R_{kj} \approx \int_0^\infty q_{Kj}(t)\gamma_k e^{-\gamma_k t}\, dt = \gamma_k \tilde{q}_{Kj}(\gamma_k),\ 0 \le j \le K,\ 1 \le k \le m, \tag{23}$$

where $q_{ij}(t)$ is the transition function of a Markov chain given in Assumption IC and $\tilde{q}_{ij}(s)$ is the Laplace transform of $q_{ij}(t)$.

Once R_{kj} is obtained, the approximations for λ_j, L_k and L_{kj} are obtained by substituting R_{kj} into (22), (18) and (20), respectively. The unknowns λ_j and R_{kj} can be calculated by fixed point iteration as **Algorithm M** in Section 2.

Let W be the time period during which a customer sojourns in the system until the customer leaves the system and R be the number of retrials made by a customer during its sojourn in the system. Let $\{R=n,\ \text{Success}\}$ be the event that a customer succeeds in getting service after the nth retrial and $\{R=n,\ \text{Loss}\}$ the event that a customer leaves the system without service after the nth retrial. Let $P_S(n) = P(R=n,\ \text{Success})$, $P_L(n) = P(R=n,\ \text{Loss})$ and

$$P_R(n) = P(R=n) = P_S(n) + P_L(n), n = 0, 1, \cdots, m.$$

The approximation formulae for $P_S(n) = P(R=n,\ \text{Success})$ and $P_L(n) = P(R=n,\ \text{Loss})$ as follows

$$\begin{aligned} P_S(n) &= \begin{cases} P_{0K}H_1(1-\gamma_1\tilde{q}_{KK}(\gamma_1)), & n=1, \\ \frac{1}{\lambda}\gamma_{n-1}L_{n-1,K}P_{0K}H_n(1-\gamma_n\tilde{q}_{KK}(\gamma_n)), & 2 \le n \le m, \end{cases} \\ P_L(n) &= \frac{1}{\lambda}\gamma_n L_{n,K}P_{0K}\bar{H}_{n+1},\ 1 \le n \le m. \end{aligned}$$

The loss probability P_L is given by

$$P_L = \sum_{n=0}^{m} P_L(n) = \frac{P_{0K}}{\lambda}\left(\lambda\bar{H}_1 + \sum_{n=1}^{m}\gamma_n L_{nK}\bar{H}_{n+1}\right).$$

Shin and Moon [56] derive an approximation formula for LST of W. Here, we present an approximation formula for $\mathbb{E}[W]$ as

$$\mathbb{E}[W] \quad = \quad (1-P_L)\frac{1}{\mu} + \sum_{n=1}^{m} P_R(n)\bar{W}_R(n)$$

$$+\sum_{i=c}^{K-1}\left(P_{0i}+\sum_{n=1}^{m}P_R(n,i)\right)\frac{i-c+1}{c\mu},$$

where $\bar{W}_R(n)=\sum_{j=1}^{n}\frac{1}{\gamma_j}$, $n=1,2,\cdots,m$ and

$$P_R(n,i)=\begin{cases} P_{0K}H_1\gamma_1\tilde{q}_{Ki}(\gamma_1), & n=1,\\ \frac{1}{\lambda}\gamma_{n-1}L_{n-1,K}P_{0K}H_n\gamma_n\tilde{q}_{Ki}(\gamma_n), & 2\le n\le m.\end{cases}$$

Numerical experiments in [56] show that approximations for $L_0=\mathbb{E}[X_0]$ are overestimated and the relative error Err(%) increases as ρ increases and γ increases, but approximations for the mean number L_{Orbit} of customers in orbit are underestimated as m increases. The behaviors of L_0 and L_{Orbit} increase, but P_L decreases as γ decreases and m increases. The performance measures L_0, L_{Orbit}, P_B and P_L as a function of m approaches to a constant as m tends to infinity.

4.3. Bibliographical notes. A retrial queueing model taking into account nonpersistence of customers was considered by [12] for the $M/M/c$ type retrial queue. Much effort has been spent to analyze the retrial queues with the special case of persistence function $H_1=\alpha$, $H_2=H_3=\cdots=\beta$. Closed form solutions for the system have not been obtained except for a few special cases, for example, $M/G/1$ retrial queue with $\alpha\le 1$ and $\beta=1$ and $M/M/1$ retrial queue with $\alpha\le 1$ and $\beta\le 1$, see [19, Section 3.3]. For multiple server case, some algorithms and approximations are presented, e.g. [19, 20, 23, 50, 61]. For more detailed references and the related results, see [19, Chapter 5], [5, Chapter 3] and [64, Chapter 7]. Generalized truncation method for the stationary distribution of $M/M/s$ retrial queue in which α and β may depend on the number of customers in service facility is presented in [50]. An approximation of the $M/M/c$ retrial queue with the persistence function $H_j=1$, $1\le j\le m$ and $H_j=0$ for $j\ge m+1$ for a given constant m is presented in [52].

5. $M/M/c$ retrial queue with multiclass of customers

5.1. The model and preliminaries. Consider an $M/M/c$ retrial queue with multiclass of customers. Customers belong to one of m different types. The customers of type i (i-customers) arrive from outside according to a Poisson process with rate λ_i, $1\le i\le m$. Note that λ_i in this section has different meaning from those in the previous sections. Let $\Lambda=\sum_{i=1}^{m}\lambda_i$ be the total arrival rate and $\alpha_i=\frac{\lambda_i}{\Lambda}$, $1\le i\le m$. When an arriving i-customer finds an available server, then the customer starts to get service and leaves the system after service. Otherwise, the customer joins orbit and repeats its request after exponential amount of time with rate γ_i. Service time distribution of i-customers is exponential with parameter μ_i. Let $\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_m)$ and $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_m)$. An approximation for this system is proposed in [59] and we describe the method and the results therein briefly in the following.

Let $C_i(t)$ be the number of i-customers being served and $N_i(t)$ the number of i-customers in orbit at time t and $\boldsymbol{C}(t) = (C_1(t), \cdots, C_m(t))$, $\boldsymbol{N}(t) = (N_1(t), \cdots, N_m(t))$. Then $\boldsymbol{X} = \{(\boldsymbol{C}(t), \boldsymbol{N}(t)),\ t \geq 0\}$ forms a Markov chain with state space $\mathcal{S} = \mathcal{K} \times \mathbb{Z}_+^m$, where $\mathcal{K} = \cup_{i=0}^c \mathcal{K}(i)$ and $\mathcal{K}(i) = \{(k_1, \cdots, k_m) \in \mathbb{Z}_+^m : \sum_{j=1}^m k_j = i\}$. The necessary and sufficient condition for the Markov chain $\boldsymbol{X}$ to be positive recurrent is $\rho = \sum_{i=1}^m \frac{\lambda_i}{c\mu_i} < 1$. Let $\boldsymbol{C} = (C_1, \cdots, C_m)$ and $\boldsymbol{N} = (N_1, \cdots, N_m)$ be the stationary version of $\boldsymbol{C}(t)$ and $\boldsymbol{N}(t)$ and $P(\boldsymbol{k}, \boldsymbol{n}) = P(\boldsymbol{C} = \boldsymbol{k}, \boldsymbol{N} = \boldsymbol{n})$, where $\boldsymbol{k} = (k_1, \cdots, k_m) \in \mathcal{K}$, $\boldsymbol{n} = (n_1, \cdots, n_m) \in \mathbb{Z}_+^m$. The balance equations for $P(\boldsymbol{k}, \boldsymbol{n})$ can be obtained by usual argument and are omitted here. It can be seen from the balance equations for $P(\boldsymbol{k}, \boldsymbol{n})$ that the following Proposition 5.1 and $\mathbb{E}[C_i] = \frac{\lambda_i}{\mu_i}$, $i = 1, 2, \cdots, m$.

Proposition 5.1. *The marginal distribution $\boldsymbol{\pi} = (\pi(\boldsymbol{k}), \boldsymbol{k} \in \mathcal{K})$ with $\pi(\boldsymbol{k}) = P(\boldsymbol{C} = \boldsymbol{k})$ is given by $\boldsymbol{\pi} Q_{MC} = 0$, where*

$$Q_{MC} = \begin{array}{c} \\ \mathcal{K}(0) \\ \mathcal{K}(1) \\ \vdots \\ \mathcal{K}(c-1) \\ \mathcal{K}(c) \end{array} \begin{array}{c} \begin{array}{ccccc} \mathcal{K}(0) & \mathcal{K}(1) & \mathcal{K}(2) & \cdots & \mathcal{K}(c) \end{array} \\ \left(\begin{array}{ccccc} D_0 & A_0 & & & \\ B_1 & D_1 & A_1 & & \\ & \ddots & \ddots & \ddots & \\ & & B_{c-1} & D_{c-1} & A_{c-1} \\ & & & B_c & D_c \end{array} \right) \end{array}. \tag{24}$$

The matrix D_l is a diagonal matrix of size $\binom{m+l-1}{m-1}$, $l = 0, 1, \cdots, c$ whose diagonal elements are determined by making $Q_{MC}\mathbf{e} = 0$ and A_l and B_l are given as follows

$$[A_l]_{\boldsymbol{jk}} = \begin{cases} a_i(\boldsymbol{j}), & \boldsymbol{k} = \boldsymbol{j} + \mathbf{e}_i,\ \boldsymbol{j} \in \mathcal{K}(l),\ \boldsymbol{k} \in \mathcal{K}(l+1),\ i = 1, 2, \cdots, m \\ 0, & \textit{otherwise} \end{cases}$$

$$[B_l]_{\boldsymbol{jk}} = \begin{cases} j_i \mu_i, & \boldsymbol{k} = \boldsymbol{j} - \mathbf{e}_i,\ \boldsymbol{j} \in \mathcal{K}(l), \boldsymbol{k} \in \mathcal{K}(l-1),\ i = 1, 2, \cdots, m \\ 0, & \textit{otherwise,} \end{cases}$$

where

$$a_i(\boldsymbol{k}) = \lambda_i + \gamma_i L_i(\boldsymbol{k}),\ i = 1, 2, \cdots, m$$

is the total arrival rate of i-customers from outside and orbit into the service facility given that the service facility is in state $\boldsymbol{k}$ and $L_i(\boldsymbol{k}) = \mathbb{E}[N_i \,|\, \boldsymbol{C} = \boldsymbol{k}]$.

It can be seen from the balance equation for $P(\boldsymbol{k}, \boldsymbol{n})$ that after tedious algebra

$$L_i = \frac{\lambda_i P_B}{\gamma_i} + \sum_{\boldsymbol{k} \in \mathcal{K}(c)} L_i(\boldsymbol{k}) \pi(\boldsymbol{k}),\ i = 1, 2, \cdots, m,$$

where $P_B = P(\boldsymbol{C} \in \mathcal{K}(c))$ is the blocking probability.

Let $R_i(\boldsymbol{k})$ be the proportion of returning customers of i-customers from orbit who find the service facility in state $\boldsymbol{k}$, that is,

$$R_i(\boldsymbol{k}) = \frac{\gamma_i \mathbb{E}[N_i 1_{\{\boldsymbol{C} = \boldsymbol{k}\}}]}{\gamma_i \mathbb{E}[N_i]} = \frac{L_i(\boldsymbol{k}) \pi(\boldsymbol{k})}{L_i},\ i = 1, 2, \cdots, m,$$

where $L_i = \mathbb{E}[N_i]$. Let $R_B(i) = \sum_{\boldsymbol{k}\in\mathcal{K}(c)} R_i(\boldsymbol{k})$. Combining the results above, we have the following proposition.

Proposition 5.2.

$$L_i = \frac{\lambda_i P_B}{\gamma_i(1 - R_B(i))}, \ i = 1, 2, \cdots, m, \tag{25}$$

$$a_i(\boldsymbol{k}) = \lambda_i + \frac{\gamma_i L_i R_i(\boldsymbol{k})}{\pi(\boldsymbol{k})}, i = 1, 2, \cdots, m. \tag{26}$$

Let $\Gamma = \sum_{i=1}^m \gamma_i$ and $\beta_i = \gamma_i/\Gamma$, $1 \le i \le m$. Fix a retrial ratio $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)$ and in order to highlight the dependence of retrial rate Γ we write the stationary distribution as $P_\Gamma(\boldsymbol{k}, \boldsymbol{n})$ instead of $P(\boldsymbol{k}, \boldsymbol{n})$.

Theorem 5.3. [59] *Assume that* $\lim_{\Gamma\to\infty} P_\Gamma(\boldsymbol{k}, \boldsymbol{n}) = P_\infty(\boldsymbol{k}, \boldsymbol{n})$ *exists. Then* $P_\infty(\boldsymbol{k}, \boldsymbol{n}) = 0$ *for* $|\boldsymbol{k}| \le c-1$, $|\boldsymbol{n}| \ge 1$ *and* $P_\infty(\boldsymbol{k}, \boldsymbol{n})$ *satisfies the balance equations for the stationary distribution in multiclass M/M/c queue with discriminatory random order service (DROS) discipline in which an i-customer is randomly selected for next service with probability* $\frac{n_i\beta_i}{\boldsymbol{\beta}\cdot\boldsymbol{n}}$ *when there are* $\boldsymbol{n} = (n_1, \cdots, n_m)$ *customers in queue upon a service completion.*

5.2. Approximations. Based on the formula Q_{MC} in (24) for $\pi(\boldsymbol{k})$, we adopt the following approximation assumption about the behavior of service facility.

Assumption MC. *The service facility behaves like a level dependent quasi-birth-and-death process with generator* Q_{MC} *and is independent of the retrials.*

Assume that an i-customer is blocked at time $t = 0$ and the customer returns after exponential time with rate γ_i. The $R_i(\boldsymbol{k})$ is approximated by the probability that this customer finds the service facility of state $\boldsymbol{k}$ upon retrial under Assumption MC, that is,

$$R_i(\boldsymbol{k}) \approx \frac{\gamma_i}{P_B} \sum_{\boldsymbol{j}\in\mathcal{K}(c)} \pi(\boldsymbol{j})[(\gamma_i I - Q_{MC})^{-1}]_{\boldsymbol{jk}}, \ i = 1, 2, \cdots, m. \tag{27}$$

Remark 5.1. If the service rates are identical, that is, $\mu_i = \mu$, $i = 1, 2, \cdots, m$, then the marginal distribution $\pi(k) = P\left(\sum_{i=1}^m C_i = k\right)$ satisfies the

$$(a(k) + k\mu)\pi(k) = a(k-1)p(k-1) + (k+1)\mu\pi(k+1), 0 \le k \le c, \tag{28}$$

where $a(k) = \Lambda + \sum_{i=1}^m \gamma_i \mathbb{E}[N_i \,|\, C = k]$ and hence we have that

$$\pi(k) = \frac{a(0)a(1)\cdots a(k-1)}{k!\mu^k}\pi(0), \ k = 1, 2, \cdots, c$$

with $\sum_{k=0}^c \pi(k) = 1$. Furthermore, the formula (27) becomes $R_i(k) = \gamma_i[(\gamma_i I - Q)^{-1}]_{ck}$, where Q is the generator corresponding to (28).

The following algorithm is for computing $\boldsymbol{a} = (\boldsymbol{a}(\boldsymbol{k}), \boldsymbol{k} \in \mathcal{K})$, where $\boldsymbol{a}(\boldsymbol{k}) = (a_1(\boldsymbol{k}), \cdots, a_m(\boldsymbol{k}))$. Write Q_{MC} as $Q_{MC}(\boldsymbol{a})$ to highlight the dependence of $\boldsymbol{a}$.

Algorithm MC.

For $n = 0, 1, 2, \cdots$, repeat the following steps starting with $a_i^{(0)}(\boldsymbol{k}) = \lambda_i$, $i = 1, \cdots, m$,

```
1. Compute the stationary distribution $\boldsymbol{\pi}^{(n)}$ of $Q^{(n)} = Q_{MC}(\boldsymbol{a}^{(n)})$
2. Compute $R_i^{(n)}(\boldsymbol{k})$ and $L_i^{(n)}$ using (27)
3. Update $a_i^{(n+1)}(\boldsymbol{k})$ with $R_i^{(n)}(\boldsymbol{k})$ and $L_i^{(n)}$ using (25) – (27)
until
```

$$||\boldsymbol{a}^{(n+1)} - \boldsymbol{a}^{(n)}|| < \epsilon.$$

5.3. Performance measures. *Blocking probability and mean number of customers.* The blocking probability is given by $P_B = \sum_{\boldsymbol{k}\in\mathcal{K}(c)} \boldsymbol{\pi}(\boldsymbol{k})$ and $\mathbb{E}[C_i] = \lambda_i/\mu_i$. Denote by $L_i(\Gamma)$ the mean number of i-customers in orbit in the system with retrial rate Γ and $L_{i,\text{App}}(\Gamma)$ the approximation of $L_i(\Gamma)$ that can be calculated by (25). In order to improve the accuracy of approximation, Shin and Moon [59] propose the modified formula $\hat{L}_i(\Gamma)$ for $L_i(\Gamma)$ as follows:

$$\hat{L}_i(\Gamma) = L_{i,\text{App}}(\Gamma) + (L_{i,M/M/c/DROS(\boldsymbol{\beta})} - L_{i,\text{App}}(\Gamma^*)),$$

where $L_{i,M/M/c/ROS(\boldsymbol{\beta})}$ is the mean number of customers in queue for ordinary $M/M/c$ queue with $DROS(\boldsymbol{\beta})$ service rule and Γ^* is sufficiently large.

Waiting time distribution. The approximation formulae for the LST of the distribution of the sojourn time W_i of an i-customer are presented in terms of the distribution of the number $N_R(i)$ of retrials made by a customer during its waiting time in the system in [59]. Here, we present the expectations as follows

$$\mathbb{E}[W_i] = \frac{1}{\gamma_i}\mathbb{E}[N_R(i)] + \frac{1}{\mu_i}, \ i = 1, 2, \cdots, m,$$

where

$$\mathbb{E}[N_R(i)] = \frac{P_B}{1 - R_B(i)}.$$

It can be seen that the approximation formula for $\mathbb{E}[W_i]$ satisfies Little's formula $\lambda_i\mathbb{E}[W_i] = \mathbb{E}[C_i] + L_i$.

Numerical results in [59] show that the approximation of P_B works well for small ρ or small Γ and becomes worse as Γ and c increase and the approximation of L_i works well for wide range of Γ. The approach for approximating L_i uses the ordinary $M/M/c/DROS$ queue. For more wide range of applications of the approach proposed here, further research is required to develop an algorithm for ordinary $M/M/c/DROS$ queue.

5.4. Bibliographical notes. There are a number of analytical results for retrial queueing models with two classes of customers that in case of blocking, one class of customers joins orbit and the other can be queued [11, 18, 28] or leaves the system without service [38]. Explicit expressions for the mean number of customers or mean waiting time in single server retrial queue with two or more classes of customers are presented in [31, 32, 16]. The $M^X/G/1$ retrial queue with multiclass customers is studied by means of branching processes with immigration in [24] and the stability condition for multiclass retrial queue with multiple servers is given in [30]. For the multiserver case with finite size of orbit, an algorithmic approach using matrix geometric method [10] or computational approache using the colored generalized Petri nets [21] are proposed. Theorem 5.3 is the version of the system with multiclass of customers corresponding to the convergence of the retrial queue with single class of customers [19, 51, 44].

6. $M/M/c$ retrial queue with server vacations

6.1. The model and preliminaries. Consider the $M/M/c$ retrial queue in which the servers take vacations. The arrival rate and service rate of each server is λ and μ, respectively. If any a $(1 \leq a < c)$ or more servers are idle at a service completion, that is, the number of customers at the service facility is less than or equal to $a^* = c - a$ upon a service completion, then b $(0 \leq b \leq a)$ servers among idle servers take a vacation and the remaining $b^* = c - b$ servers are available. This vacation policy is called the (a, b)-vacation policy [65]. The vacation time distribution is assumed to be of a phase type $PH(\boldsymbol{\delta}, \boldsymbol{V})$, where $\boldsymbol{V} = (v_{ij})$ is a nonsingular $w \times w$ matrix with $v_{ii} = -v_i < 0$, $1 \leq i \leq w$ and $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_w)$ with $\boldsymbol{\delta}\mathbf{e} = 1$. Let $\boldsymbol{V}^0 = -\boldsymbol{V}\mathbf{e} = (v_1^0, \cdots, v_w^0)^T$ and $m_v = \boldsymbol{\delta}(-\boldsymbol{U})^{-1}\mathbf{e}$ be the mean vacation time. We consider the single vacation policy under which the servers take only one vacation and after the vacation the servers either serves the waiting customer in service facility if any or stays idle.

If a customer finds that the number of customers in service facility is less than c upon arrival, the customer enters the service facility, otherwise the customer joins orbit and repeats its request until the customer gets into the service facility. The customers in orbit retry independently with other customers and retrial times of each customer are assumed to be independent and identically distributed. We assume that the retrial time distribution of a customer in orbit is of phase type $PH(\boldsymbol{\theta}, \boldsymbol{U})$ whose distribution function is $F(t) = 1 - \boldsymbol{\theta} \exp(\boldsymbol{U}t)\mathbf{e}$, $t \geq 0$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_g) \geq 0$ with $\boldsymbol{\theta}\mathbf{e} = 1$ and $\boldsymbol{U} = (u_{ij})$ is a nonsingular $g \times g$ matrix with $u_{ii} = -u_i < 0$, $1 \leq i \leq g$. Let $\boldsymbol{u} = (u_1, \cdots, u_g)$, $\boldsymbol{\gamma} = -\boldsymbol{U}\mathbf{e} = (\gamma_1, \cdots, \gamma_g)^T$ and $m_r = \boldsymbol{\theta}(-\boldsymbol{U})^{-1}\mathbf{e}$ be the mean retrial time. Here, we introduce the results in [60] and the readers can refer the paper for details.

Let $X_i(t)$ the number of customers in orbit whose service phase is of i, $1 \leq i \leq g$ and $Y(t)$ be the number of customers at service facility and $J(t)$ the server state at time t defined by

$$J(t) = \begin{cases} 0, & c \text{ servers are available} \\ j, & \text{the phase of vacation time is of } j,\ 1 \leq j \leq w. \end{cases}$$

Then $\boldsymbol{\Psi} = \{(\boldsymbol{X}(t), Y(t), J(t)), t \geq 0\}$ with $\boldsymbol{X}(t) = (X_1(t), \cdots, X_g(t))$ is a continuous time Markov chain on the state space $\mathcal{S} = \{(\boldsymbol{n}, k, j) \in \mathbb{Z}_+^{g+2} : \boldsymbol{n} \geq 0, 0 \leq k \leq c, 0 \leq j \leq w\}$ and $\boldsymbol{\Psi}$ is positive recurrent if $\rho = \frac{\lambda}{c\mu} < 1$ [49]. Let $(\boldsymbol{X}, Y, J)$ be the stationary version of $\boldsymbol{\Psi}$. One can easily obtain the balance equations for $P(\boldsymbol{n}, k, j) = P(\boldsymbol{X} = \boldsymbol{n}, Y = k, J = j)$ and we omit the results, see [60] for details.

Let $\mathcal{Y} = \{(k,j) : 0 \le k \le c,\, 0 \le j \le w\}$ and $\gamma(k,j)$ be retrial rate from orbit given that $(Y,J) = (k,j)$, that is,

$$\gamma(k,j) = \sum_{i=1}^{g} \gamma_i L_i(k,j),\ (k,j) \in \mathcal{Y},$$

where $L_i(k,j) = \mathbb{E}[X_i | Y = k, J = j]$.

Proposition 6.1. *The marginal distribution* $\boldsymbol{\pi} = (\pi(k,j), (k,j) \in \mathcal{Y})$ *with* $\pi(k,j) = P(Y = k, J = j)$ *satisfies* $\boldsymbol{\pi} Q_V = 0$, *where*

$$Q_V = \begin{pmatrix} B_0 & A_0 & & & \\ C_1 & B_1 & A_1 & & \\ & \ddots & \ddots & \ddots & \\ & & C_{c-1} & B_{c-1} & A_{c-1} \\ & & & C_c & B_c \end{pmatrix}.$$

The matrix A_k *is the diagonal matrix of size* $w+1$ *whose diagonal elements are*

$$[A_k]_{jj} = \lambda + \gamma(k,j),\ j = 0, 1, \cdots, w,\ 0 \le k \le c-1.$$

The matrices B_k *and* C_k *are square matrices of size* $w+1$ *whose* (i,j)*-component are as follows:*

$$[B_k]_{ij} = \begin{cases} v_i^0, & 1 \le i \le w, j = 0 \\ v_{ij}, & 1 \le i \ne j \le w \\ -\Delta_k(i), & i = j, \end{cases}$$

$$[C_k]_{ij} = \begin{cases} \mu_k \delta_j, & 0 \le k \le a^* + 1, i = 0, 1 \le j \le w \\ \mu_k, & a^* + 2 \le k \le c, i = j = 0 \\ \mu_k^*, & 1 \le i = j \le w, \end{cases}$$

where $\Delta_k(i)$ *is the positive number that makes* $Q_V \mathbf{e} = 0$ *and the components not stated above are all zero and the components not stated above are all zero.*

Proposition 6.2. *Let* $\boldsymbol{L} = (L_1, \cdots, L_g)$ *with* $L_i = \mathbb{E}[X_i]$. *Then*

$$\boldsymbol{L} = \Lambda \boldsymbol{\theta} (-\boldsymbol{U})^{-1}, \tag{29}$$

where

$$\Lambda = \left(\lambda + \sum_{i=1}^{g} \gamma_i L_i(c) \right) P_B$$

and $P_B = P(Y = c)$ *and* $L_i(c) = \mathbb{E}[X_i | Y = c]$.

Note from (29) and $\boldsymbol{\theta}(-\boldsymbol{U})^{-1}\boldsymbol{\gamma} = 1$ that $\sum_{i=1}^{g} \gamma_i L_i = \Lambda$ and

$$L = \sum_{i=1}^{g} L_i = \Lambda m_r.$$

Proposition 6.3. *Let $R(k,j)$ be the proportion of returning customers from orbit who find the arrival phase and service facility in state (k,h), that is,*

$$R(k,j) = \frac{1}{\Lambda}\gamma(k,j)\pi(k,j), \;\; (k,j) \in \mathcal{Y}. \tag{30}$$

Then the mean number of customers in orbit L is given by

$$L = \frac{\lambda P_B}{1 - R_B} m_r, \tag{31}$$

where R_B is the portion of blocking of a returning customer

$$R_B = \sum_{j=0}^{w} R(c,j) = 1 - \frac{\lambda P_B}{\Lambda}.$$

6.2. Approximations. Based on the formula of Q_V, we adopt the following approximation assumption about the behavior of service facility.

Assumption V. *The service facility behaves like a level dependent quasi-birth-and-death process with generator Q_V and is independent of the retrials.*

We approximate $R(k,j)$ with the probability that a customer who joins orbit at time 0 finds the service facility is in state (k,j) at the retrial instant as follows:

$$R(k,j) \approx \frac{1}{\Lambda}\sum_{i=0}^{w}(\lambda + \gamma(c,i))\pi(c,i)\left[\int_0^\infty e^{Q_V t}\boldsymbol{\theta} e^{\boldsymbol{U}t} dt\right]_{(c,i),(k,j)}. \tag{32}$$

Once initial value of $\gamma(k,j)$ is given, $\Lambda R(k,j)$ can be approximated by (32) using the stationary distribution $\boldsymbol{\pi}$ and $\gamma(k,j)$ is updated from $\Lambda R(k,j)$ by the formula (30). The following algorithm summarizes the results above. We write Q_V as $Q_V(\gamma)$ to highlight the dependence of γ.

Algorithm V.

```
For $n = 0, 1, 2, \cdots$, repeat with $\gamma^{(0)}(k,j) = 0$
    1. Let $Q^{(n)} = Q_V(\gamma^{(n)})$ and compute $\boldsymbol{\pi}^{(n)}$;
    2. Compute $\Lambda R^{(n)}(k,j)$ using (32);
    3. Update $\gamma^{(n+1)}(k,j)$ using (30);
until
    $||\gamma^{(n+1)} - \gamma^{(n)}|| < \epsilon.$
```

6.3. Performance measures. Once $\boldsymbol{\pi}$ and $R(k,j)$ are obtained through Algorithm V, the performance measures such as the blocking probability $P_B = P(Y = c)$, the probability $P_V = 1 - P(J = 0)$ that the servers are in vacation, the mean $\mathbb{E}[Y]$ and standard deviation $SD[Y]$ of the number of customers in service facility can be calculated. The mean number $L(m_r)$ of customers in orbit when the mean retrial time is m_r is approximated by the formula

$$\hat{L}(m_r) = L_{\text{App}}(m_r) + (L_V - L_{\text{Appr}}(m_r^*)), \tag{33}$$

where $L_{\text{App}}(m_r)$ is the approximation formula (31) and L_V is the mean number of customers waiting in the queue for the ordinary $M/M/c$ vacation queue. The quantity m_r^* is chosen to be small enough so that the variation of $L_{\text{App}}(m_r)$ is negligible for $m_r \leq m_r^*$. It follows from the numerical examples in [60] that the accuracy of approximation is good in practical sense and tends to improve as the mean retrial time m_r increases.

6.4. Bibliographical notes. Retrial queues and vacation queues have been studied separately for last several decades. Recently, the interests on the retrial queues with vacations is growing rapidly. However, almost all the literature deals with the system with single-server and/or constant retrial policy that only one customer in orbit can retry e.g. see [1, 8, 27, 14, 34]. Algorithmic approaches for the single server queue with Bernoulli vacation schedule and linear retrial policy are considered by [13]. The call center with outgoing calls introduced in [43] can be considered as the queues with retrials and vacations. An algorithmic solution for the $MAP/M/c/K$ queue with PH-vacation time and exponential retrial time is developed in [48] and the stability condition for the $MAP/PH/c/K$ queue with PH vacation time and PH retrial time is given in [49].

7. Concluding remarks

The behavior of the retrial queue is described by multidimensional process jointly describing the states of service facility and the states of orbit. The common features of the method described above are as follows.

The first step is to reduce the whole system equations that consists of infinite number of equations in general to the system of equations of finite number equations for the marginal distributions of service facility. The reduced equations describe the behavior of service facility and contain the unknown parameters that reflect the retrials from orbit. In this step, one can see that the arrival rate to the service facility is sum of the rate from external arrivals and the rate of orbit that is unknown. Then, derive the equations for approximate the arrival rates from orbit to the service facility under appropriate approximation assumption. Finally, solve the equations for unknown parameters by iteration and calculate the performance measures.

In this paper, we have explained the dimension reduction method in detail for $M/M/c$ retrial queue and showed that the method is very useful to approximate the systems with complex structures in retrial queueing models. Besides the systems mentioned above, the method seems to be a promising approach to approximate the more complex systems.

References

1. J.R. Artalejo, *Analysis of an $M/G/1$ queue with constant repeated attempts and server vacations*, Computers & Operations Research **24** (1997), 493-504.

2. J.R. Artalejo, *A classified bibliography of research on retrial queues: Progress in 1990-1999*, **7** (1999), 187-211.

3. J.R. Artalejo, *Accessible bibliography on retrial queues*, Mathematical and Computer Modelling **30** (1999), 1-6.
4. J.R. Artalejo, *Accessible bibliography on retrial queues: Progress in 2000-2009*, Mathematical and Computer Modelling **51** (2000) 1071-1081.
5. J.R. Artalejo, A. Gómez-Corral, *Retrial Queueing Systems, A Computational Approach*, Springer-Verlag, Hidelberg, 2008.
6. S. Asmussen, *Applied Probability and Queues*, 2nd Ed., Springer-Verlag, New York, 2003.
7. A. Bobbio, A. Horváth and M. Telek, *Matching three moments with minimal acyclic phase type distributions,* Stochastic Models **21** (2005), 303-326.
8. M. Boualem, N. Djellab and D. Aïssani, *Stochastic inequality for $M/G/1$ retrial queues with vacations and constant retrial policy,* Mathematical and Computer Modelling **50** (2009), 207-212.
9. L. Breuer, A. Dudin and V. Klimenok, *A Retrial $BMAP/PH/N$ system,* Queueing Systems **40** (2002), 433-457.
10. B.D. Choi and Y. Chang, *Single server retrial queues with probability calls,* Mathematical and Computer Modelling **30** (1999), 7-32.
11. B.D. Choi, K.B. Choi and Y.W. Lee, *$M/G/1$ retrial queueing systems with two types of calls and finite capacity,* Queueing Systems **19** (1995), 215-229.
12. J.W. Cohen, *Basic problems of telephone traffic theory and the influence of repeated calls,* Philips Telecommunication Review **18** (1957), 49-100.
13. G. Choudhury, *Steady state analysis of an $M/G/1$ queue with linear retrial policy and two phase service under Bernoulli vacation schedule,* Applied Mathematical Modelling **32** (2008), 2480-2489.
14. G. Choudhry and J.C. Ke, *A batch arrival retrial queue with general retrial times under Bernoulli vacation schedule for unreliable server and delayed repair,* Applied Mathematical Modelling **36** (2012), 255-269.
15. J.E. Diamond, A.S. Alfa, *Approximation method for $M/PH/1$ retrial queues with phase type inter-retrial times,* European Journal of Operational Research **113** (1999), 620-631.
16. G.I. Falin, *On a multiclass batch arrival retrial queue,* Advances in Applied Probability **20** (1988), 483-487.
17. G.I. Falin, *A survey of retrial queues,* Queueing Systems **7** (1990), 127-168.
18. G.I. Falin, J.R. Artalejo and M. Martin, *On the single server retrial queue with priority customers,* Queueing Systems **14** (1993), 439-455.
19. G.I. Falin and J.G.C. Templeton, Retrial Queues, Chapman and Hall, London, 1997.
20. A.A. Fredericks and G.A. Reisner, *Approximations to stochastic service systems with an application to a retrial model,* Bell Systems Technical Journal **58** (1979), 557-576.
21. N. Gharbi, C. Dutheillet and M. Ioualalen, *Colored stochastic Petri nets for modelling and analysis of multiclass retrial systems,* Mathematical and Computer Modelling **49** (2009), 1436-1448.
22. A. Gómez-Corral, *A bibliographical guid to the analysis of retrial queues through matrix analytic techniques,* Annals of Operations Research **141** (2006), 163-191.
23. B.S. Greenberg and R.W. Wolf, *An upper bound on the performance of queues with returning customer,* Journal of Applied Probbaility **24** (1987), 466-475.
24. S.A. Grishechkin, *Multiclass batch arrival retrial queues analyzed as branching process with immigration,* Queueing Systems **11** (1992), 395-418.
25. Q.M. He, H. Li, Y.Q. Zhao, *Ergodicity of the $BMAP/PH/s/s+K$ retrial queue with $PH-$retrial times,* Queueing Systems **35** (2000), 323-347.
26. M.A. Johnson, M.R. Taaffe, *Matching moments to phase distributions : mixture of Erlang distributions of common order,* Stochastic Models **5** (1989), 711-743.
27. J.C. Ke and F.M. Chang, *Modified vacation policy for $M/G/1$ retrial queue with balking and feedback,* Computers & Industrial Engineering **57** (2009), 433-443.

28. Z. Khalil, G.I. Falin and T. Yang, *Some analytical results for congestion in subscriber line modules,* Queueing Systems **10** (1992), 381-402.
29. J. Kim and B. Kim, *A survey of retrial queueing systems,* Annals of Operations Research **247** (2016), 3-36.
30. J. Kim, J. Kim J and B. Kim, *Analysis of the $M/G/1$ queue with discriminatory random order service policy,* Performance Evaluation **68** (2011), 256-270.
31. V.G. Kulkarni, *On queueing systems with retrials,* Journal of Applied Probability **20** (1983), 380-389.
32. V.G. Kulkarni, *Expected waiting times in a multiclass batch arrival retrial queue,* Journal of Applied Probability **23** (1986), 144-154.
33. V.G. Kulkarni and H.M. Liang, *Retrial queues revisited, In Frontiers in Queueing: Models and Applications in Science and Engineering* (J.H. Dshalalow, ed.), CRC Press, Boca Raton, 19-34.
34. B.K. Kummar, R. Rukmani and V. Thangaraj, *An $M/M/c$ retrial queueing system with Bernoulli vacations,* Journal of Systems Science and Systems Engineering **18** (2009), 222-242.
35. H.M. Liang, *Retrial queues*, Ph,D Thesis, University of North Carolina at Chapel Hill, 1991.
36. H.M. Liang, *Service station factors in monotonicity of retrial queues,* Mathematical and Computer Models **30** (1999), 189-196.
37. H.M. Liang, V.G. Kulkarni, *Monotonicity properties of single server retrial queues,* Stochastic Models **9** (1993), 373-400.
38. M. Martin and J.R. Artalejo, *Analysis of an $M/G/1$ queue with two types of impatient units,* Advances in Applied Probability **27** (1995), 840-861.
39. E. Morozov, *A multiserver retrial queue: regenerative stability analysis,* Queueing Systems **56** (2007), 157-168.
40. M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach, Johns Hopkins University Press, Baltimore, 1981.
41. M.F. Neuts and B.M. Rao, *Numerical investigation of a multiserver retrial model,* Queueing Systems **7** (1990), 169-190.
42. T. Osogami and M. Harchol-Balter, *Closed form solutions for mapping general distributions to quasiminimal PH distributions,* Performance Evaluation **62** (2006), 524-552.
43. T. Phung-Duc and K. Kawanishi, *An efficient method for performance analysis of blended call centers with retrial,* Asia Pacific Journal of Operational Research **31** (2014), 1440008 (33pages).
44. Y.W. Shin, *Monotonicity properties in various retrial queues and their applications,* Queueing Systems **53** (2006), 147-157.
45. Y.W. Shin, *Fundamental matrix of transient QBD generator with finite states and level dependent transitions,* Asia-Pacific Journal of Operational Research **26** (2009), 697-714.
46. Y.W. Shin, *Algorithmic solutions for $M/M/c$ retrial queue with PH_2 retrial times,* Journal of Applied Mathematics and Informatics **29** (2011), 803-811.
47. Y.W. Shin, *Interpolation approximation of $M/G/c/K$ retrial queues with ordinary queues,* Journal of Applied Mathematics and Informatics **30** (2012), 531-540.
48. Y.W. Shin, *Algorithmic approach to Markovian multi-server retrial queue with vacations,* Applied Mathematics and Computations **250** (2015), 287-279.
49. Y.W. Shin, *Stability of $MAP/PH/c/K$ retrial queue with customer retrials and server vacations,* Bulletin of the Korean Mathematical Society **53** (2016), 985-1004.
50. Y.W. Shin, T.S. Choo, *$M/M/s$ queue with impatient customers and retrials,* Applied Mathematical Modelling **33** (2009), 2596-2606.
51. Y.W. Shin and Y.C. Kim, *Stochastic comparisons of Markovian retrial queues,* Journal of the Korean Statistical Society **29** (2000), 473-488.

52. Y.W. Shin and D.H. Moon, *Approximations of retrial queu with limited number of retrials,* Computers and Operations Research **37** (2010), 1262-1270.
53. Y.W. Shin and D.H. Moon, *Approximation of $M/M/c$ retrial queu with PH-retrial times,* European Journal of Operational Research *213* (2011), 205-209.
54. Y.W. Shin and D.H. Moon, *Sensitivity of $M/M/c$ retrial queu with respect to retrial times : experimental investigation,* Journal of the Korean Institute of Industrial Engineers **37** (2011), 83-87.
55. Y.W. Shin and D.H. Moon, *Approximation of $M/G/c$ retrial queue with $M/PH/c$ retrial queue,* Communications of the Korean Stistical Society **19** (2012), 169-175.
56. Y.W. Shin and D.H. Moon, *Approximation of $M/M/s/K$ retrial queu with nonpersistent customers,* Applied Mathematical Modelling **37** (2013), 753-761.
57. Y.W. Shin and D.H. Moon, *On approximations for $GI/G/c$ retrial queues,* Journal of Applied Mathematics and Informatics **31** (2013), 311-325.
58. Y.W. Shin and D.H. Moon, *Approximation of $PH/PH/c$ retrial queu with PH-retrial times,* Asia-Pacific Journal of Operational Research **31** (2014), 1440010 (21 pages).
59. Y.W. Shin and D.H. Moon, *$M/M/c$ retrial queue with multiclass of customers,* Merhodology and Computing in Applied Probability **16** (2014), 931-949.
60. Y.W. Shin and D.H. Moon, *Approximate analysis of $M/M/c$ retrial queue with server vacations,* Journal of the Korean Society for Industrial and Applied Mathematics **19** (2015), 443-457.
61. S.N. Stepanov, *Generalized model with repeated calls in case of extreme load,* Queueing Systems **27** (1997), 131-151.
62. H. Tijms, A First Course in Stochastic Models, Wiley, 2003.
63. W. Whitt, *Approximating a point process by a renewal process*, I: two basic methods, Operations Research **30** (1982), 125-147.
64. R.W. Wolff, Stochastic Modeling and The Theory of Queues, New Jersey, Prentice Hall, 1989.
65. X. Xu and Z.G. Zhang, *Analysis of multiple-server queue with a single vacation (e,d)-policy*, Performance Evaluation **63** (2006), 825-838.
66. T. Yang and J.G.C. Templeton, *A survey on retrial queues,* Queueing Systems **2** (1987), 201-233.
67. T. Yang, M.J.M. Posner, J.G.C. Templeton and H. Li, *An approximation method for the $M/G/1$ retrial queues with general retrial times,* European Journal of Operational Research **76** (1994), 552-562.

Yang Woo Shin received B.S. from Kyungpook National University, and M.Sc. and Ph.D in Mathematics at KAIST. He is currently a professor at Changwon National University. His research interests include queueing theory and its applications.

Department of Statistics, Changwon National University, Changwon, Gyeongnam 51140, Korea
e-mail:ywshin@changwon.ac.kr