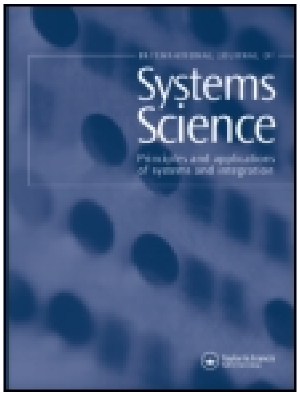

International Journal of Systems Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tsys20

Multi-server retrial queue with second optional service: algorithmic computation and optimisation

Jau-Chuan Ke [a], Chia-Huang Wu [b] & Wen Lea Pearn [b]

[a] Department of Applied Statistics, National Taichung Institute of Technology, No. 129, Sec. 3, Sanmin Road, Taichung, Taiwan 404, ROC

[b] Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu, Taiwan 300, ROC

Published online: 13 Dec 2010.

To cite this article: Jau-Chuan Ke , Chia-Huang Wu & Wen Lea Pearn (2011) Multi-server retrial queue with second optional service: algorithmic computation and optimisation, International Journal of Systems Science, 42:10, 1755-1769, DOI: 10.1080/00207721003650209

To link to this article: http://dx.doi.org/10.1080/00207721003650209

International Journal of Systems Science
Vol. 42, No. 10, October 2011, 1755–1769

Multi-server retrial queue with second optional service: algorithmic computation and optimisation

Jau-Chuan Ke[a]*, Chia-Huang Wu[b] and Wen Lea Pearn[b]

[a]*Department of Applied Statistics, National Taichung Institute of Technology, No. 129, Sec. 3, Sanmin Road, Taichung, Taiwan 404, ROC;* [b]*Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu, Taiwan 300, ROC*

(*Received 23 March 2009; final version received 4 January 2010*)

We consider an infinite-capacity M/M/c retrial queue with second optional service (SOS) channel. An arriving customer finds a free server would enter the service (namely, the first essential service, denoted by FES) immediately; otherwise, the customer enters into an orbit and starts generating requests for service in an exponentially distributed time interval until he finds a free server and begins receiving service. After the completion of FES, only some of them receive SOS. The retrial system is modelled by a quasi-birth-and-death process and some system performance measures are derived. The useful formulae for computing the rate matrix and stationary probabilities are derived by means of a matrix-analytic approach. A cost model is derived to determine the optimal values of the number of servers and the two different service rates simultaneously at the minimal total expected cost per unit time. Illustrative numerical examples demonstrate the optimisation approach as well as the effect of various parameters on system performance measures.

Keywords: cost; first essential channel; quasi-Newton method; second optional channel; matrix-geometric method

1. Introduction

This article considers an M/M/c retrial queue in which primary customers arrive to a Poisson process with parameter λ. An arriving primary customer finding one or more servers available (free) obtains service immediately. On the other hand, if the primary customer finds all servers busy, he joins the orbit and tries to get the service later on. There are c channels (servers) that provide the first essential service (FES) as well as the second optional service (SOS) to arriving customers. The FES is needed by all arriving customers. The service times of the FES and the SOS have an exponential distribution with means $1/\mu_1$ and $1/\mu_2$, respectively. As soon as the FES of a customer is completed, a customer may leave the system with probability $(1-\theta)$ or may opt for the SOS with probability θ $(0 \le \theta \le 1)$, at the completion of which the customer departs from the system and the next customer, if any, from the queue is taken up for his FES (Figure 1). Each channel can serve only one customer at a time and it also provides only one essential service or SOS at a time. Furthermore, each customer staying in the orbit makes repeated attempts, in random intervals having length exponentially distributed with parameter σ, independently of the other customers. Upon requesting service from the orbit, a customer who finds all c servers busy always rejoins the orbit; this manner continues until he is eventually served. We assume that there exists an upper bound N on the number of customers in the orbit that are allowed to conduct retrials (Neuts and Rao 1990; Artalejo and Pozo 2002). This implies that the probability of a repeated attempt during $(t, t+\mathrm{d}t)$, given that j customers in the orbit at time t, is $\sigma_j \mathrm{d}t + o(\mathrm{d}t)$, where $\sigma_j = \min\{j, N\}\sigma$. Moreover, we assume that the process of primary arrivals, service times and inter-retrial times are mutually independent.

An arbitrary customer in the orbit generates a stream of repeated requests that is independent of the rest of the customers in the orbit. This situation arises in telephony, where an arriving call is not allowed to await the termination of a busy signal. Such queueing systems play important roles in the analysis of many telephone switching systems, telecommunication networks and computer systems. The following are just a few examples of problems that can be modelled as retrial queues with SOS.

(1) Telephone systems: A subscriber who obtains a busy signal will repeat the request until the connection is made. Therefore, the flow of calls

*Corresponding author. Email: jauchuan@ntit.edu.tw

ISSN 0020–7721 print/ISSN 1464–5319 online

DOI: 10.1080/00207721003650209
http://www.informaworld.com

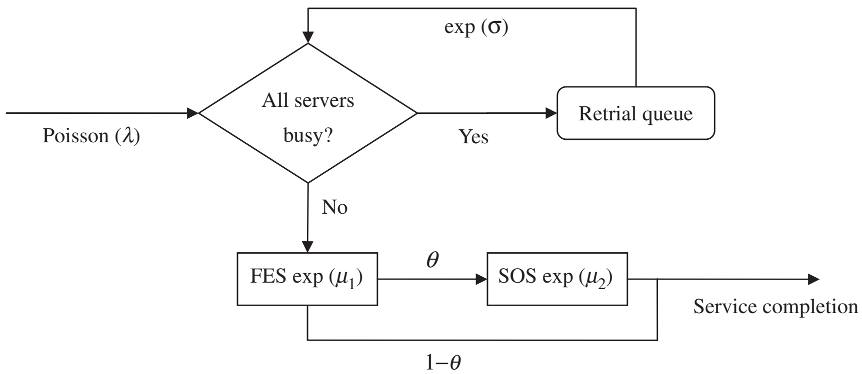

Figure 1. The general structure of M/M/c retrial queue with second optional service.

circulating in a telephone network consists of the flow of primary calls and the flow of repeated calls. The former reflects the real wishes of the telephone subscribers and the latter is the retrial request of previous attempts. When a subscriber makes a successful phone call (connect to switchboard), the call duration will consist of an FES that this subscriber who is phoning asks for the person that he is looking for and then it is determined whether such an interlocutor is present or not (based on Artalejo and Choudhury (2004)). Moreover, in the booking ticket service, a subscriber who demands a booking ticket service by telephone can address his request (FES) and receive the proper fundamental service (SOS) if the stock of tickets is still not empty. This telephone system (usually with multi-servers) can be regarded as a practical application of our model.

(2) Communication network: An Internet network with multi-servers is very common in daily life. A customer who finds all servers busy will try to connect again after a space. The servers provide browse service (FES) for each entering user and supply SOS such as online shopping, network information or object download services (SOS) for some of them.

(3) Daily life queues: We often observe that a person finding a long waiting line chooses to balk and return later. This situation can be regarded as the retrial phenomenon due to balking behaviour where the retrial is motivated by impatience. In a cafeteria, everyone needs to fill a bowl with rice and take some well-cooked foods (FES), but some of them prefer to order more freshly made edibles (SOS). Furthermore, for example, a driver who arrives at a gas station and sees a long waiting line leaves to handle other daily task or job and returns to gas station as back home. All cars arriving at a gas station may need gas refuelling services (FES), but only some of them may require a car wash services (SOS) after refuelling.

Reviews of retrial queue literature can be found in Yang and Templeton (1987), Falin and Templeton (1997) and Artalejo (1999a, b). A number of applications of retrial queues in science and engineering can be found in Kulkarni and Liang (1997). Apart from its practical interest due to its more accurate representation of several congestion phenomena, the multi-server retrial queue raises interesting mathematical and computational questions. The investigation of the multi-server retrial queues is essentially more difficult than single-server models. Explicit formulae for the stationary distribution of a M/M/c retrial queue are known only when the number of servers c is at most two. Most multi-server retrial queues can be modelled by a level-dependent quasi-birth-and-death (QBD) process. The main feature of its infinitesimal generator is the spatial heterogeneity caused by transitions due to repeated attempts. This lack of homogeneity causes the analytical complexity of retrial models. Many interesting studies have been devoted to an approximate approach to the stationary probabilities for system states (Falin 1983; Neuts and Rao 1990; Bright and Taylor 1995; Stepanov 1999; Artalejo and Lopez-Herrero 2000; Artalejo and Pozo 2002; Breuer, Dudin, and Klimenok 2002; Chakravarthy and Dudin 2002). Recently, Gomez-Corral (2006) gave a detailed bibliographical guide to the analysis of retrial queues through matrix analytic techniques.

The truncation models seem to be the most convenient method for obtaining reliable numerical solutions for the M/M/c retrial queue. For example, Falin (1983) assumed that the retrial rate becomes infinite when the number of customers in orbit exceeds a level M. It means that, from the level M up, the system performs as an ordinary M/M/1 queue with arrival rate λ and service rate $c\mu$ so that the condition $\lambda < c\mu$ is necessary and sufficient for the ergodicity. Neuts and Rao (1990) and Artalejo and Pozo (2002) proposed several models in this direction and approximated efficiently the stationary distribution of the M/M/c retrial queue. As related works, a number of studies investigated the computation of the other system characteristics, such as the distributions of busy period, successful and blocked retrials, for the multi-server retrial queue of type M/M/c. Readers can refer to Artalejo, Chakravarthy, and Lopez-Herrero (2007a), Artalejo, Economou, and Lopez-Herrero (2007b), Amador and Artalejo (2007) and others. Artalejo, Economou, and Lopez-Herrero (2007c) presented an algorithmic analysis of the maximum number of customers in orbit (and in the system) during a busy period for the M/M/c retrial queue. The multi-server retrial queueing problems are extensively studied as mentioned above. However, there is no work on a multi-server retrial queue with a SOS channel in the literature. In this article, we investigate the multi-server retrial queue with a SOS channel via a matrix-geometric approach.

Studies on various queueing models in the past are characterised by a common feature; all customers receive service in the first phase (so-called main service) by the channel (server), and they do not further request other SOS. However, in many real service systems, one encounters numerous examples of the queueing situations where all arrivals require the main service and only some of them may request the subsidiary service provided by the channel. Analytic steady-state solutions of an M/M/c retrial queue with second optional channel have not been found. A pioneering work in this queueing situation is proposed by Madan (2000), who first introduced the concept of SOS (channel). Madan (2000) studied an M/G/1 queue with SOS, using the supplementary variable technique in which he considered general service time distribution for FES and exponential service time distribution for SOS. He also cited some important applications of this model in many real-life situations. Madan (2001) extended Madan's model (2000) to two-stage service channel systems with generally distributed. Medhi (2002) derived the transient solution and steady-state solution for the ordinary M/G/1 queue with SOS using the same technique. Medhi's M/G/1 model was also investigated by Al-Jararha and Madan (2003), in which they developed the time-dependent probability generating functions involved in Laplace transforms and further obtained the corresponding steady-state results. The reliability measures were examined by Wang (2004), for the ordinary M/G/1 queue with SOS and considering channel breakdowns. Ke (2008) investigated a batch arrival $M^{[x]}$/G/1 queueing system with J optional services. He derived the steady-state results, including system size distribution at a random epoch and at a departure epoch, the distributions of idle and busy periods and waiting time distribution in the queue. Recently, Choudhury and Tadj (2009) generalised this type of model by introducing the concept of a server breakdown and a delay-repair period. The optimal control and management of such models have also received considerable attention in the literature. For example, Choudhury and Madan (2005) investigated such a type of model for a two-stage batch arrival queue with Bernoulli vacation schedule and Choudhury and Paul (2006) studied a similar type of model for a batch arrival queueing system with two phases of service. They derived the queue size distribution at a random epoch as well as at a departure epoch for an $M^{[x]}$/G/1 queueing system with second optional channel under N-policy. A simple procedure was also provided to obtain optimal stationary policy under a suitable linear cost structure. Tadj, Choudhury, and Tadj (2006a, b) investigated some bulk service queueing system under N-policy. As for the retrial models with related works, Artalejo and Choudhury (2004) examined the steady state behaviour of an M/G/1 queue with repeated attempts in which the server may provide an additional second phase of service. Some practical applications were presented in Artalejo and Choudhury's works which also generalised both the classical M/G/1 retrial queue and the M/G/1 queue with classical waiting line and SOS. An M/G/1 retrial queueing system with two phases of service subject to the server breakdown and repair was investigated by Choudhury and Deka (2008), who derived the queue size distribution at a random epoch and departure epoch using supplementary variable technique, various system performance measures were also presented. Ke and Chang (2009) investigated a batch-arrival M/G/1 retrial queue under Bernoulli vacation schedules with two phases of service general repeated attempts and starting failures.

Existing retrial queueing problems with optional service, including the above mentioned, mainly focussed on *one single-server queue*. Because of the more complicated structure of the stochastic processes required to describe the system behaviours, the multi-server retrial queue is known to be analytically intractable. This motivates us to investigate a multi-server retrial queue of M/M/c type with second optional channel.

The article is organised as follows. In Section 2, the QBD model of the multi-server M/M/c retrial queue with SOS is set up. The rate matrix and stable condition of the QBD model are derived using matrix-geometric property. In Section 3, an efficient algorithm is developed to calculate the stationary probabilities by matrix-geometric method. In Section 4, some system performance measures are derived. In Section 5, a cost model is developed to determine the optimal number of servers and two service rates, in order to minimise the total expected cost per unit time. We use a direct search method and quasi-Newton method to implement the optimisation tasks. Some numerical examples are provided to illustrate these two optimisation methods. In Section 6, conclusions are made with some remarks.

2. M/M/c retrial queue with SOS

Consider an M/M/c retrial queue with SOS. The state of the system can be described as (i, j, k), where i and k denote the number of servers busy in the FES and SOS, respectively. j is the number of customers in orbit (sources of repeated demands). The system can be described by a continuous parameter Markov chain on the state space $\{(i, j, k);\ 0 \le i \le c,\ 0 \le j,\ 0 \le k \le c - i\}$. From Figure 2, the customers entering the server get services immediately as $i + k < c$ (i.e. there are available servers). The new arriving customer who finds all c servers busy $(i + k = c)$ always rejoins the retrial group (orbit), this operation continues until they are eventually served. In the steady state, we define $P^k_{i,j} \equiv$ probability that there are i and k servers busy in the FES and SOS, respectively, and j customers in orbit, where $0 \le i + k \le c$, $j \ge 0$.

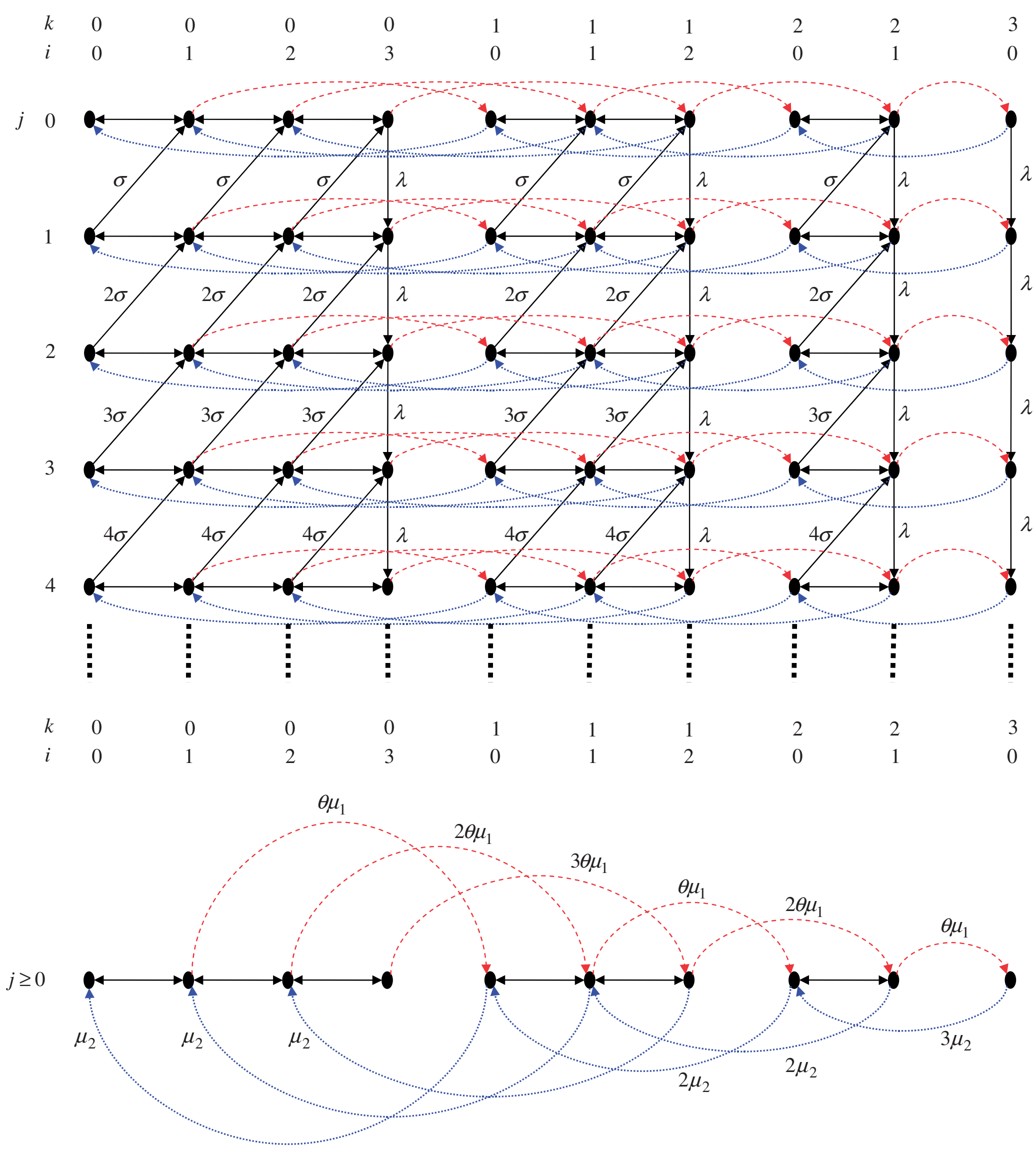

Figure 2. Steady-transition-rate diagram for a multi-server retrial queue with second optional service ($c = 3$).

2.1. *Matrix representation*

The infinitesimal generator Q of the Markov chain has the form

$$Q=\begin{bmatrix} A_0 & B & & & & & & & \\ C_1 & A_1 & B & & & & & & \\ & C_2 & A_2 & B & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & C_N & A_{N-1} & B & & & \\ & & & & C_N & A_N & B & & \\ & & & & & C_N & A_N & B & \\ & & & & & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{1}$$

The infinitesimal generator Q of that Markov chain has the form shown in Figure 3 for $c=3$. The entries B, $A_i(i\geq 0)$, and $C_i(i\geq 1)$ are square matrices of order $(c+1)(c+2)/2$. Block-diagonal square matrices B and C_i can be partitioned as:

$$B=\begin{bmatrix} b^0 & & & & \\ & b^1 & & & \\ & & \ddots & & \\ & & & b^{c-1} & \\ & & & & b^c \end{bmatrix} \quad \text{and}$$

$$C_i=\begin{bmatrix} c_i^0 & & & & \\ & c_i^1 & & & \\ & & \ddots & & \\ & & & c_i^{c-1} & \\ & & & & c_i^c \end{bmatrix}, \quad i=1,2,\ldots$$

$k=0$				$k=1$			$k=2$		$k=3$	$k=0$				$k=1$			$k=2$		$k=3$
$-\lambda$	λ			0						0									
$q\mu_1$	$*$	λ		$\theta\mu_1$							0								
	$2q\mu_1$	$*$	λ		$2\theta\mu_1$							0							
		$3q\mu_1$	$*$			$3\theta\mu_1$							λ						
μ_2				$*$	λ		0							0					
	μ_2			$q\mu_1$	$*$	λ	$\theta\mu_1$								0				
		μ_2	0		$2q\mu_1$	$*$		$2\theta\mu_1$								λ			
				$2\mu_2$			$*$	λ	0								0		
					$2\mu_2$	0	$q\mu_1$	$*$	$\theta\mu_1$									λ	
							$3\mu_2$	0	$*$										λ
0	σ									$*$	λ								
	0	σ								$q\mu_1$	$*$	λ							
		0	σ								$2q\mu_1$	$*$	λ						
			0									$3q\mu_1$	$*$						
				0	σ					μ_2				$*$	λ				
					0	σ					μ_2			$q\mu_1$	$*$	λ			
						0						μ_2	0		$2q\mu_1$	$*$			
							0	σ						$2\mu_2$			$*$	λ	
								0							$2\mu_2$	0	$q\mu_1$	$*$	
									0								$3\mu_2$	0	$*$
										0	2σ								
											0	2σ							
												0	2σ						
													0						
														0	2σ				
															0	2σ			
																0			
																	0	2σ	
																		0	
																			0

Figure 3. The infinitesimal generator Q ($c=3$, $q=1-\theta$).

where sub-matrices b^j and c_i^j are square matrices of order $(c+1-j)$ with elements

$$\begin{cases} b^j[c+1-j, c+1-j] & = \lambda \\ 0 & e.w \end{cases} \quad \text{and}$$

$$\begin{cases} c_i^j[k, k+1] = \sigma_i = \min\{i, N\}\sigma, & 1 \le k \le c-j \\ 0 & e.w \end{cases}.$$

A_i can be partitioned as

$$A_i = \begin{bmatrix} Y_i^0 & X^0 & & & & & \\ Z^1 & Y_i^1 & X^1 & & & & \\ & Z^2 & Y_i^2 & X^2 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & Z^{c-1} & Y_i^{c-1} & X^{c-1} \\ & & & & & Z^c & Y_i^c \end{bmatrix},$$

$$i = 0, 1, 2, \ldots$$

where X^j is a $(c+1-j)\times(c-j)$ matrix with $X^j[k+1, k] = k\theta\mu_1$, $1 \le k \le c-j$, Z^j is a $(c-j)\times(c+1-j)$ matrix with $Z^j[k,k] = j\mu_2$, $1 \le k \le c-j$ and Y_i^j is a square matrix of order $(c+1-j)$ with elements

$$\begin{cases} Y_i^j[k, k+1] = \lambda, \quad 1 \le k \le c-j \\ Y_i^j[k+1, k] = k(1-\theta)\mu_1, \quad 1 \le k \le c-j \\ Y_i^j[1,1] = -[\lambda + j\mu_2 + \sigma_i] \\ Y_i^j[k,k] = -[\lambda + (k-1)\mu_1 + j\mu_2 + \sigma_i], \quad 2 \le k \le c-j \\ Y_i^j[c+1-j, c+1-j] = -[\lambda + j\mu_2 + (c-j)\mu_1] \end{cases}.$$

For instance, for $c = 3$, the sub-matrices of B are

$$b^0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix}, \quad b^1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \lambda \end{bmatrix},$$

$$b^2 = \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix}, \quad b^3 = \lambda.$$

The sub-matrices of C_1 are

$$c_1^0 = \begin{bmatrix} 0 & \sigma & 0 & 0 \\ 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & \sigma \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad c_1^1 = \begin{bmatrix} 0 & \sigma & 0 \\ 0 & 0 & \sigma \\ 0 & 0 & 0 \end{bmatrix},$$

$$c_1^2 = \begin{bmatrix} 0 & \sigma \\ 0 & 0 \end{bmatrix}, \quad c_1^3 = 0.$$

For A_1, the first super diagonal sub-matrices are

$$X^0 = \begin{bmatrix} 0 & 0 & 0 \\ \theta\mu_1 & 0 & 0 \\ 0 & 2\theta\mu_1 & 0 \\ 0 & 0 & 3\theta\mu_1 \end{bmatrix}, \quad X^1 = \begin{bmatrix} 0 & 0 \\ \theta\mu_1 & 0 \\ 0 & 2\theta\mu_1 \end{bmatrix},$$

$$X^2 = \begin{bmatrix} 0 \\ \theta\mu_1 \end{bmatrix}.$$

The diagonal sub-matrices are

$$Y_1^0 = \begin{bmatrix} -(\lambda+\sigma) & \lambda & & \\ (1-\theta)\mu_1 & -(\lambda+\mu_1+\sigma) & \lambda & \\ & 2(1-\theta)\mu_1 & -(\lambda+2\mu_1+\sigma) & \lambda \\ & & 3(1-\theta)\mu_1 & -(\lambda+3\mu_1) \end{bmatrix},$$

$$Y_1^1 = \begin{bmatrix} -(\lambda+\mu_2+\sigma) & \lambda & \\ (1-\theta)\mu_1 & -(\lambda+\mu_1+\mu_2+\sigma) & \lambda \\ & 2(1-\theta)\mu_1 & -(\lambda+2\mu_1+\mu_2) \end{bmatrix},$$

$$Y_1^2 = \begin{bmatrix} -(\lambda+2\mu_2+\sigma) & \lambda \\ (1-\theta)\mu_1 & -(\lambda+\mu_1+2\mu_2) \end{bmatrix},$$

$$Y_1^3 = -(\lambda+3\mu_2).$$

The first sub diagonal sub-matrices are

$$Z_1 = \begin{bmatrix} \mu_2 & 0 & 0 & 0 \\ 0 & \mu_2 & 0 & 0 \\ 0 & 0 & \mu_2 & 0 \end{bmatrix}, \quad Z_2 = \begin{bmatrix} 2\mu_2 & 0 & 0 \\ 0 & 2\mu_2 & 0 \end{bmatrix},$$

$$Z_3 = \begin{bmatrix} 3\mu_2 & 0 \end{bmatrix}.$$

2.2. *Rate matrix R*

Let $\Pi = [\Pi_0, \Pi_1, \Pi_2, \ldots]$ with $\Pi_i = [P_{0,i}^0, P_{1,i}^0, \ldots, P_{c,i}^0, P_{0,i}^1, P_{1,i}^1, \ldots, P_{c-1,i}^1, \ldots, P_{0,i}^{c-1}, P_{1,i}^{c-1}, P_{0,i}^c]$, $\quad i = 0, 1, 2, \ldots$ be the unique solution to $\Pi Q = \mathbf{0}$ and $\Pi e = 1$, where e is a column vector with all elements equal to 1. It is noted that the vector $\Pi = [\Pi_0, \Pi_1, \Pi_2, \Pi_3, \ldots]$ has the following properties:

$$\Pi_{N+k} = \Pi_N R^k, \quad \text{for } k \ge 1. \tag{2}$$

The matrix R is the unique non-negative solution with spectral radius less than one of the equations:

$$B + RA_N + R^2 C_N = \mathbf{0}. \tag{3}$$

From Neuts (1981) and Latouche and Ramaswami (1999), we know R is given by $\lim_{n\to\infty} R_n$, where the sequence $\{R_n\}$ is defined by

$$R_0 = \mathbf{0}, \text{ and } R_{n+1} = -BA_N^{-1} - R_n^2 C_N A_N^{-1}, \quad \text{for } n \ge 0. \tag{4}$$

The sequence $\{R_n\}$ is monotone so that R could be evaluated from (4) by successive substitutions.

2.3. *Stability condition*

It is known (Theorem 3.1.1 of Neuts 1981) that the steady-state probability vector exists if and only if

$$xBe < xC_N e, \tag{5}$$

where x is the invariant probability of the matrix $F = C_N + A_N + B$. x satisfies $xF = \mathbf{0}$ and $xe = 1$ where e is a column vector with dimension $(c+1)(c+2)/2$ and all elements equal to one. Substituting B and C_N into Equation (5) and doing some routine manipulation leads to

$$N\sigma(1 - P_F) > \lambda P_F, \tag{6}$$

where P_F denotes the probability that all servers are busy (i.e. $i+k=c$), i.e. the system will be stable if the expected successful retrial rate is greater than the expected arrival rate of 'orbit'.

3. Steady-state solution

Under the stability condition, the stationary probability vector Π of Q exists. In this section, we deal with the steady-state equations by representing it in matrix form. This steady-state probability vector $\Pi = [\Pi_0, \Pi_1, \Pi_2, \Pi_3, \ldots]$ is given by

$$\Pi_0 A_0 + \Pi_1 C_1 = \mathbf{0}, \tag{7a}$$

$$\Pi_{i-1}B + \Pi_i A_i + \Pi_{i+1}C_{i+1} = \mathbf{0}, \quad 1 \le i \le N-1, \tag{7b}$$

$$\Pi_{N-1}B + \Pi_N A_N + \Pi_N R C_N = \mathbf{0}, \tag{7c}$$

$$\Pi_N R^{i-1-N}B + \Pi_N R^{i-N}A_N + \Pi_N R^{i+1-N}C_N = \mathbf{0}, \quad N+1 \le i, \tag{7d}$$

$$\sum_{i=0}^{\infty} \Pi_i e = 1. \tag{8}$$

After doing some routine manipulations to Equation (7a)–(7c), we have

$$\begin{aligned}\Pi_0 &= \Pi_1 C_1(-A_0)^{-1} = \Pi_1\phi_1,\\ \Pi_{i-1} &= \Pi_i C_i[-(\phi_{i-1}B + A_{i-1})]^{-1} = \Pi_i\phi_i, \quad 2 \le i \le N,\end{aligned} \tag{9}$$

and

$$\Pi_N\phi_N B + \Pi_N A_N + \Pi_N R C_N = \mathbf{0}. \tag{10}$$

Consequently, Π_i $(0 \le i \le N-1)$ in Equation (9) can be written as product form in terms of Π_N and the rest steady-state vector $[\Pi_N, \Pi_{N+1}, \Pi_{N+2}, \ldots]$ can be determined recursively as $\Pi_i = \Pi_N R^{i-N}$, for $i \ge N$. Once the steady-state probability Π_N is obtained, the steady-state solutions $[\Pi_0, \Pi_1, \Pi_2, \ldots, \Pi_{N-1}, \Pi_N, \Pi_{N+1}, \ldots]$ are determined. The steady-state probability Π_N can be solved by Equation (10) with the following normalisation equation:

$$\begin{aligned}\sum_{i=0}^{\infty}\Pi_i e &= [\Pi_0 + \Pi_1 + \cdots + \Pi_{N-1} + \Pi_N + \Pi_{N+1}\\ &\quad + \Pi_{N+2} + \cdots]e\\ &= \Bigg[\Pi_N \prod_{i=N}^{1}\phi_i + \Pi_N\prod_{i=N}^{2}\phi_i + \cdots + \Pi_N\prod_{i=N}^{N}\phi_i\\ &\quad + \Pi_N + \Pi_N R + \Pi_N R^2 + \cdots\Bigg]e\\ &= \Pi_N\left[\sum_{k=1}^{N}\prod_{i=N}^{k}\phi_i + (I-R)^{-1}\right]e = 1.\end{aligned} \tag{11}$$

Solving Equations (10) and (11) in accordance with Cramer's rule, we obtain Π_N. Then, the prior state probabilities $[\Pi_0, \Pi_1, \Pi_2, \ldots, \Pi_{N-1}]$ are computed from (9) and $[\Pi_{N+1}, \Pi_{N+2}, \Pi_{N+3}, \ldots]$ are gained by the formula $\Pi_i = \Pi_N R^{i-N}$, $i \ge N+1$. We summarise the solution procedure of steady-state probabilities as below:

Algorithm: Recursive solver

Step 1: Set $\phi_1 = C_1(-A_0)^{-1}$

Step 2: For i from 2 to N, set $\phi_i = C_i[-(\phi_{i-1}B + A_{i-1})]^{-1}$.

Step 3: For k from 1 to N, set $\Phi_k = \prod_{i=N}^{k}\phi_i$.

Step 4: Solve $\Pi_N\phi_N B + \Pi_N A_N + \Pi_N R C_N = \mathbf{0}$, $\Pi_N[\sum_{k=1}^{N}\Phi_k + (I-R)^{-1}]e = 1$ and obtain steady-state probability Π_N.

Step 5: Construct steady-state probability Π_i as follows:

(1) if $0 \le i \le N$, assign $\Pi_i = \Pi_N\Phi_{i+1}$,
(2) if $N+1 \le i$, assign $\Pi_{i+1} = \Pi_i R$,

4. System performance measures

The system performance measures, such as the expected number of customers in the FES channel (denoted by $E[FES]$), the expected number of customers in the SOS channel (denoted by $E[SOS]$) and the expected number of customers in orbit (denoted by $E[Orbit]$), can be evaluated from the steady-state probabilities $\Pi_i = [P^0_{0,i}, P^0_{1,i}, \ldots, P^0_{c,i}, P^1_{0,i}, P^1_{1,i}, \ldots, P^1_{c-1,i}, \ldots, P^{c-1}_{0,i}, P^{c-1}_{1,i}, P^c_{0,i}]$. The expressions for

$E[FES]$, $E[SOS]$ and $E[Orbit]$ are given by

$$\begin{aligned} E[FES] &= \sum_{i=0}^{\infty} \Pi_i v = \sum_{i=0}^{N-1} \Pi_i v + \Pi_N v + \Pi_N R v \\ &\quad + \Pi_N R^2 v + \cdots \\ &= \sum_{i=0}^{N-1} \Pi_N \Phi_{i+1} v + \Pi_N v + \Pi_N R v + \Pi_N R^2 v + \cdots \\ &= \Pi_N \left[\sum_{i=1}^{N} \Phi_i + (I-R)^{-1} \right] v, \end{aligned} \tag{12}$$

$$\begin{aligned} E[SOS] &= \sum_{i=0}^{\infty} \Pi_i J = \sum_{i=0}^{N-1} \Pi_i J + \Pi_N J + \Pi_N R J \\ &\quad + \Pi_N R^2 J + \cdots \\ &= \sum_{i=0}^{N-1} \Pi_N \Phi_{i+1} J + \Pi_N (I-R)^{-1} J \\ &= \Pi_N \left[\sum_{i=1}^{N} \Phi_i + (I-R)^{-1} \right] J, \end{aligned} \tag{13}$$

$$\begin{aligned} E[Orbit] &= \sum_{i=1}^{\infty} i \Pi_i e = \sum_{i=1}^{N-1} i \Pi_N \Phi_{i+1} e + N \Pi_N e \\ &\quad + (N+1) \Pi_N R e + (N+2) \Pi_N R^2 e + \cdots \\ &= \sum_{i=2}^{N} (i-1) \Pi_N \Phi_i e + \Pi_N [N(I-R)^{-1} \\ &\quad + R(I-R)^{-2}] e \\ &= \Pi_N \left[\sum_{i=2}^{N} (i-1) \Phi_i + N(I-R)^{-1} + R(I-R)^{-2} \right] e, \end{aligned} \tag{14}$$

where

$$v = [\underbrace{0, 1, \ldots, c}_{\#=c+1}, \underbrace{0, 1, \ldots, c-1}_{\#=c}, \ldots, \underbrace{0, 1}_{\#=2}, 0]$$

and

$$J = [\underbrace{0, 0, \ldots, 0}_{\#=c+1}, \underbrace{1, 1, \ldots, 1}_{\#=c}, \ldots, \underbrace{c-1, c-1}_{\#=2}, c]$$

are column vectors with dimension $(c+1)(c+2)/2$.

For an M/M/c retrial queue with SOS channel, the numerical results of $E[Orbit]$ are obtained by considering the following three cases with different values of c:

Case 1: $N=30$, $\lambda=5$, $\mu_2=10$, $\theta=0.5$, $\sigma=5$, vary μ_1 from 10 to 20.

Case 2: $N=30$, $\lambda=5$, $\mu_1=10$, $\theta=0.5$, $\sigma=5$, vary μ_2 from 10 to 20.

Case 3: $N=30$, $\mu_1=20$, $\mu_2=15$, $\theta=0.5$, $\sigma=5$, vary λ from 5 to 10.

The results of $E[Orbit]$ are depicted in Figure 4 for Case 1–3, respectively. From Figure 4, it should be noted that $E[Orbit]$ is insensitive to the change of μ_1, μ_2 and λ when the number of servers is greater than one.

There are several general descriptors of retrial queues, some of which are listed as follows:

(1) The overall rate of retrials:

$$\begin{aligned} \sigma_1^* &= \sum_{j=1}^{N} j\sigma \sum_{k=0}^{c} \sum_{i=0}^{c-k} P_{i,j}^k + \sum_{j=N+1}^{\infty} N\sigma \sum_{k=0}^{c} \sum_{i=0}^{c-k} P_{i,j}^k \\ &= \sum_{j=1}^{N} j\sigma \Pi_j e + \sum_{j=N+1}^{\infty} N\sigma \Pi_N R^{j-N} e \\ &= \sum_{j=1}^{N} j\sigma \Pi_j e + N\sigma \Pi_N R (I-R)^{-1} e \\ &= \sigma \left[\sum_{j=1}^{N} j \Pi_j + N \Pi_N R (I-R)^{-1} \right] e \\ &= \sigma \Pi_N \left[\sum_{j=1}^{N-1} j \Phi_{j+1} + N(I-R)^{-1} \right] e. \end{aligned} \tag{15}$$

(2) The rate of retrials that are successful:

$$\sigma_2^* = \sum_{j=1}^{N} j\sigma \sum_{k=0}^{c} \sum_{i=0}^{c-k-1} P_{i,j}^k + \sum_{j=N+1}^{\infty} N\sigma \sum_{k=0}^{c} \sum_{i=0}^{c-k-1} P_{i,j}^k. \tag{16}$$

(3) The fraction of successful retrials:

$$F = \frac{\sigma_2^*}{\sigma_1^*}. \tag{17}$$

(4) The marginal distribution of the number of busy servers:

$$\sum_{j=0}^{\infty} P_{i,j}^k, \quad 0 \le i+k \le c. \tag{18}$$

(5) Busy period: The busy period T of a retrial queue is defined as the period that starts at the epoch when an arriving customer finds an empty system (all servers are idle and no customer in the orbit) and ends at the departure epoch at which the system is empty again.

The mean busy period

$$E(T) = \frac{1}{\lambda} \left(\frac{1}{P_{0,0}^0} - 1 \right) = \frac{1}{\lambda} \left(\frac{1}{\Pi_N \Phi_1[1]} - 1 \right), \tag{19}$$

where $\Pi_N \Phi_1[1]$ denotes the first element of $\Pi_N \Phi_1$.

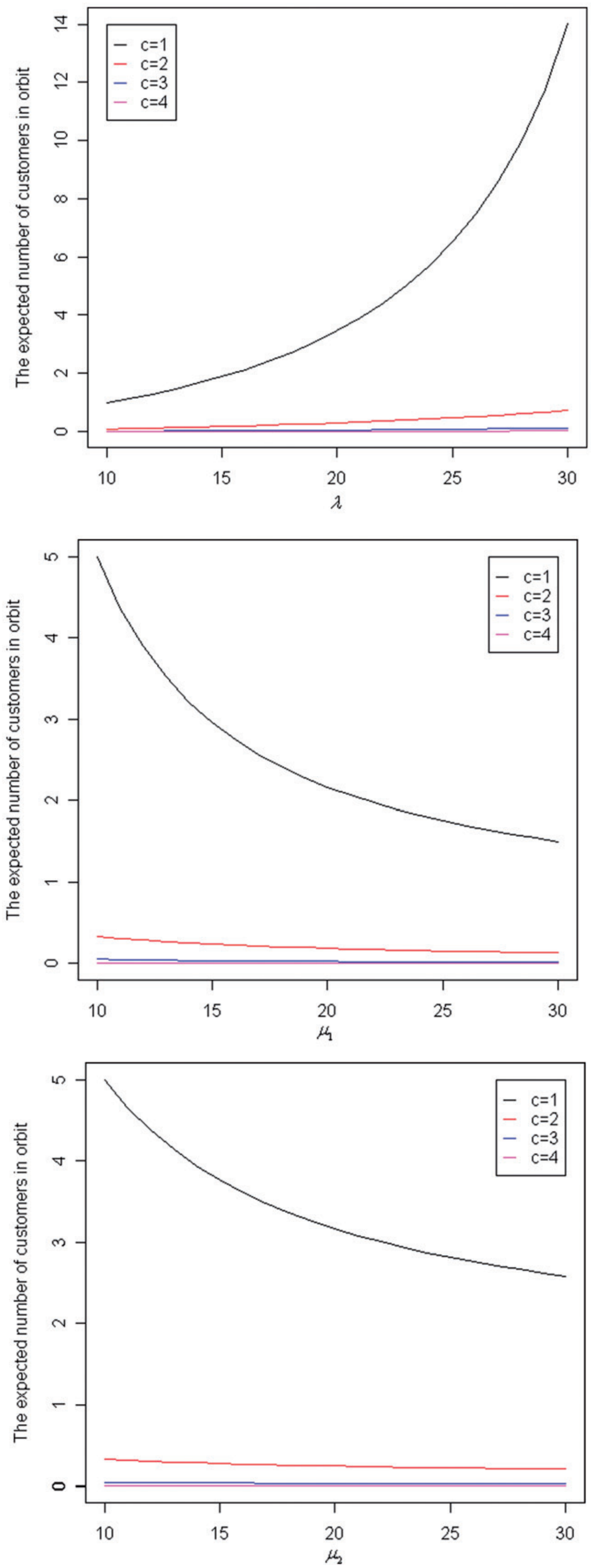

Figure 4. The expected number of customers in orbit *vs.* λ, μ_1 and μ_2.

(6) Vain retrials: A vain retrial is an unsuccessful retrial when all servers are busy.

The steady-state conditional probability of vain retrial P_v is defined as (Krishna and Raja 2006; Krishna, Rukmani, and Thangaraj 2009)

$$P_V = \frac{\sum_{j=1}^{\infty} \sum_{i+k=c} P_{i,j}^k}{\sum_{j=1}^{\infty} \sum_{k=0}^{c} \sum_{i=0}^{c-k} P_{i,j}^k} = \frac{\sum_{j=1}^{\infty} \sum_{i+k=c} P_{i,j}^k}{1 - \Pi_0 e}. \quad (20)$$

To understand how the system performance measures listed above vary with N, we also perform a numerical investigation to the measures based on changing the value of N. The numerical illustration is graphically presented in Figure 5.

From Figure 5, it is clear that increasing the retrial rate beyond a certain point does not result in a commensurate improvement in the system performance, i.e. when the number of customers in orbit is sufficiently large, a majority of the retrial requests fail to find a free server and do not result in a change of state. Therefore, the number of customers who can generate retrial requests could be restricted (truncated) to an appropriately chosen number N (Neuts and Rao 1990).

5. Optimisation analysis

We construct a total expected cost function per unit time, in which the number of servers (c) is a discrete decision variable, and the service rates $\mu = (\mu_1, \mu_2)$ are continuous decision variables. Let us define the following cost elements:

- C_h cost per unit time per customer present in orbit,
- C_1 cost per unit time when one server is busy,
- C_2 cost per unit time of providing a service rate μ_1,
- C_3 cost per unit time of providing a service rate μ_2,
- C_4 fixed cost to purchase one server.

Based on the definition of the cost parameters, the total expected cost function per unit time is given by

$$\begin{aligned} F(c, \mu_1, \mu_2) = C_h E[Orbit] + C_1(E[FES] \\ + E[SOS]) + C_2\mu_1 + C_3\mu_2 + C_4 c. \quad (21) \end{aligned}$$

The main objective is to determine the optimal number of servers $c*$, and the optimal value of the service rate $\mu* = (\mu_1^*, \mu_2^*)$, simultaneously which minimise the cost function. The analytic study of the optimisation behaviour of the expected cost function would have been an arduous task to undertake

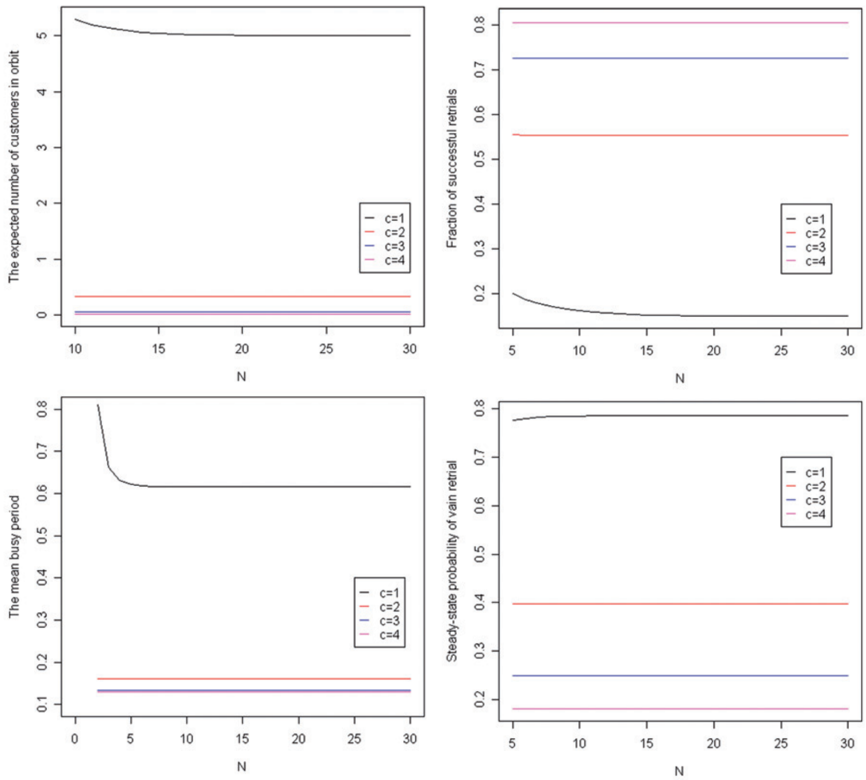

Figure 5. The system performance measures *vs.* the truncated parameter N.

since the decision variable appears in an expression which is a highly complex and non-linear in terms of (c, μ_1, μ_2).

We first use direct search method to find the optimal value of the number of servers, say c^*, when μ_1 and μ_2 are fixed. Next, we fix c^* and use the quasi-Newton method to search/adjust the optimal value of (μ_1, μ_2), say (μ_1^*, μ_2^*).

5.1. *Direct search method*

For practical application, an upper bound U is imposed on c. We can successively substitute $c = 1, 2, \ldots, U$ into the cost function. The optimum value c^* could be determined by satisfying the following inequality

$$F(c^* - 1|\mu_1, \mu_2) > F(c^*|\mu_1, \mu_2) < F(c^* + 1|\mu_1, \mu_2). \qquad (22)$$

It is noted that $F(c^*|\mu_1, \mu_2)$ is a (local) minimum. To demonstrate that the cost function is really convex in c and the solution gives a minimum, some numerical examples are performed based on the preceding formulation. For convenience, the number $N = 30$ is chosen and the following cost elements are adopted:

$$C_h = \$25/\text{customer/unit time},$$
$$C_1 = \$120/\text{server/unit time}, \; C_2 = \$15/\text{unit time},$$
$$C_3 = \$30/\text{unit time}, \; C_4 = \$180/\text{server}.$$

Under other parameters that are given, we observe from Table 1 that the optimal number of servers c^* and its corresponding minimum cost increase as θ or λ increases and decrease as σ increases.

5.2. *Quasi-Newton method*

After we obtain c^*, a quasi-Newton method is employed to search (μ_1, μ_2) until the minimum value of $F(c^*, \mu_1, \mu_2)$ is achieved, say $F(c^*, \mu_1^*, \mu_2^*)$. To find the joint optimal value (μ_1^*, μ_2^*) for a given c^*, we should show the convexity of $F(c^*, \mu_1, \mu_2)$. However, this work is difficult to implement. We note that the derivative of the cost function F with respect to $\mu = (\mu_1, \mu_2)$ indicates the direction in which cost function increases. It means that the optimal value (μ_1^*, μ_2^*) can be found along this opposite direction of the gradient (Chong and Zak 2001).

Table 1. The cost function associated with number of servers and values of λ.

$(\mu_1, \mu_1, \lambda, \theta, \sigma)$	$c=1$	$c=2$	$c=3$	$c=4$	$c=5$	$c=6$
(20, 10, 5, 0.2, 5)	**840.75**	1003.57	1182.15	1362.01	1542.00	1722.00
(20, 10, 10, 0.2, 5)	**1026.54**	1056.44	1225.94	1404.29	1584.04	1764.00
(20, 10, 15, 0.2, 5)	N/A[a]	**1133.07**	1274.45	1447.72	1626.32	1806.05
(20, 10, 20, 0.2, 5)	N/A	**1277.00**	1332.65	1493.81	1669.37	1848.30
(20, 10, 5, 0.2, 10)	**834.019**	1002.95	1182.086	1362.01	1542.00	1722.00
(20, 10, 10, 0.2, 10)	**968.171**	1051.67	1225.14	1404.17	1584.02	1764.00
(20, 10, 15, 0.2, 10)	N/A	**1115.42**	1271.07	1446.99	1626.18	1806.03
(20, 10, 20, 0.2, 10)	N/A	**1222.10**	1323.03	1491.42	1668.78	1848.16
(20, 10, 5, 0.2, 15)	**831.776**	1002.74	1182.07	1362.01	1542.00	1722.00
(20, 10, 10, 0.2, 15)	**948.724**	1050.06	1224.87	1404.12	1584.02	1764.00
(20, 10, 15, 0.2, 15)	N/A	**1109.48**	1269.93	1446.75	1626.13	1806.02
(20, 10, 20, 0.2, 15)	N/A	**1203.68**	1319.78	1490.61	1668.58	1848.12
(20, 10, 5, 0.8, 5)	**930.322**	1043.87	1218.86	1398.12	1578.02	1758.00
(20, 10, 10, 0.8, 5)	N/A	**1179.97**	1306.96	1478.47	1656.55	1836.11
(20, 10, 15, 0.8, 5)	N/A	4365.50	**1431.25**	1567.86	1737.90	1915.07
(20, 10, 20, 0.8, 5)	N/A	N/A	1778.63	**1683.62**	1827.33	1996.97
(20, 10, 5, 0.8, 10)	**907.107**	1041.82	1218.53	1398.07	1578.01	1758.00
(20, 10, 10, 0.8, 10)	N/A	**1158.61**	1303.03	1477.52	1656.32	1836.06
(20, 10, 15, 0.8, 10)	N/A	2668.80	**1411.62**	1562.80	1736.40	1914.64
(20, 10, 20, 0.8, 10)	N/A	N/A	1668.51	**1665.55**	1821.65	1995.03
(20, 10, 5, 0.8, 15)	**899.369**	1041.12	1218.42	1398.06	1578.01	1758.00
(20, 10, 10, 0.8, 15)	N/A	**1151.44**	1301.70	1477.20	1656.25	1836.05
(20, 10, 15, 0.8, 15)	N/A	2347.02	**1405.00**	1561.08	1735.88	1914.49
(20, 10, 20, 0.8, 15)	N/A	N/A	**1632.44**	1659.44	1819.73	1994.38

Note: [a]Denotes system is unstable (i.e., the stable condition does not hold).

An effective procedure that makes it possible to calculate the optimal value (μ_1^*, μ_2^*) is presented as follows:

Algorithm: Quasi-Newton method

Step 1: Set the initial trial solution for $\vec{\eta}^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)})$ and compute $F(c, \mu_1^{(0)}, \mu_2^{(0)})$.

Step 2: For $i = 0, 1, 2, \ldots$, compute the cost gradient $\vec{\nabla} F(\vec{\eta}) = [\partial F/\partial\mu_1, \partial F/\partial\mu_2]^{\mathrm{T}}$ and the cost Hessian matrix $H(\vec{\eta}) = \begin{bmatrix} \partial^2 F/\partial\mu_1^2 & \partial^2 F/\partial\mu_1\partial\mu_2 \\ \partial^2 F/\partial\mu_2\partial\mu_1 & \partial^2 F/\partial\mu_2^2 \end{bmatrix}$ at point $\vec{\eta}^{(i)} = (\mu_1^{(i)}, \mu_2^{(i)})$.

Step 3: While $|\partial F/\partial\mu_1| > \varepsilon$ or $|\partial F/\partial\mu_2| > \varepsilon$ (ε denotes the tolerance, a sufficient small number), set the new trial solution $\vec{\eta}^{(i+1)} = \vec{\eta}^{(i)} - [H(\vec{\eta}^{(i)})]^{-1}\vec{\nabla} F(\vec{\eta}^{(i)})$ and return to Step 2 until the gradient is sufficiently small.

For this purpose, we present two examples to illustrate the optimisation procedure shown in Table 2. From Table 2, we can see that the minimum expected cost per unit time of 1003.92 is achieved at $(\mu_1^*, \mu_2^*) = (23.4453, 8.02222)$ by using five iterations, which is $c^* = 1$ based on Case (i) with initial value $(\mu_1, \mu_2) = (20, 10)$. Based on Case (ii), c^* is 4 and the minimum expected cost per day of 1674.11 is achieved at $(\mu_1^*, \mu_2^*) = (16.8630, 10.7441)$ by using 5 iterations.

We now perform a sensitivity investigation to the optimal value (c^*, μ_1^*, μ_2^*) based on the changes in specific values of the system parameters. The numerical results are shown in Table 3 for various values of λ, θ and σ by considering the initial value (μ_1, μ_2) of (20, 10). From Table 3, we find that (1) c^* increases as $\lambda(\theta)$ increases and is insensitive to the change of σ and (2) μ_1^* increases as λ increases and decreases as θ (or σ) increases. We also observe that (1) μ_2^* has the same pattern with μ_1^* and (2) the minimum expected cost value increases as λ (or θ) increases and decreases as σ increases.

6. Conclusions

A multi-server retrial queue with SOS was investigated using the matrix geometric method. The sufficient and necessary conditions for the stability of the system were discussed. The stationary probability vectors and the system performance measures were obtained in matrix forms. A cost model was constructed to calculate the optimal values of the number of servers

Table 2. The illustration of the implementation process of quasi-Newton method.

Iterations	0	1	2	3	4	5
Case (i): $(\lambda, \theta, \sigma) = (10, 0.2, 5)$ with initial value $(\mu_1, \mu_2) = (20, 10)$						
$F(c^*, \mu_1, \mu_2)$	1026.54	1007.83	1004.06	1003.92	1003.92	**1003.92**
c^*	1	1	1	1	1	1
μ_1	20	22.6527	23.3038	23.4405	23.4453	**23.4453**
μ_2	10	7.55324	7.92219	8.01767	8.02221	**8.02222**
$\frac{\partial F}{\partial \mu_1}$	−8.24475	−4.34979	−0.69766	−0.25822	−0.00004	-4.0×10^{-10}
$\frac{\partial F}{\partial \mu_2}$	9.22640	−10.6671	−1.80336	−0.07392	−0.00014	-1.5×10^{-9}
E[Orbit]	6.50153	7.06782	6.20192	6.02626	6.01906	6.01905
E[SB][a]	0.70000	0.70623	0.68157	0.67606	0.67583	0.67583
Hessian	$\begin{bmatrix} 5.8322 & 2.9535 \\ 2.9535 & 6.9730 \end{bmatrix}$	$\begin{bmatrix} 4.0699 & 4.6075 \\ 4.6075 & 20.781 \end{bmatrix}$	$\begin{bmatrix} 2.9736 & 3.0487 \\ 3.0487 & 14.521 \end{bmatrix}$	$\begin{bmatrix} 2.7867 & 2.7845 \\ 2.7845 & 13.382 \end{bmatrix}$	$\begin{bmatrix} 2.7796 & 2.7738 \\ 2.7738 & 13.334 \end{bmatrix}$	$\begin{bmatrix} 2.7805 & 2.7742 \\ 2.7742 & 13.334 \end{bmatrix}$
Case (ii): $(\lambda, \theta, \sigma) = (20, 0.8, 5)$ with initial value $(\mu_1, \mu_2) = (20, 10)$						
$F(c^*, \mu_1, \mu_2)$	1683.62	1675.80	1674.15	1674.11	1674.11	**1674.11**
c^*	4	4	4	4	4	4
μ_1	20	15.7834	16.6998	16.8593	16.8630	**16.8630**
μ_2	10	10.7383	10.7446	10.7442	10.7441	**10.7441**
$\frac{\partial F}{\partial \mu_1}$	4.88607	−3.28908	−0.42979	−0.00954	−0.000005	-1.3×10^{-9}
$\frac{\partial F}{\partial \mu_2}$	−2.97931	−2.06125	−0.25997	−0.00563	−0.000005	-1.0×10^{-9}
E[Orbit]	2.06473	2.64154	2.35613	2.31366	2.31270	2.31269
E[SB]	2.60000	2.75714	2.68674	2.67547	2.67521	2.67521
Hessian	$\begin{bmatrix} 1.3534 & 1.1115 \\ 1.1115 & 10.383 \end{bmatrix}$	$\begin{bmatrix} 3.5744 & 2.1813 \\ 2.1813 & 9.9203 \end{bmatrix}$	$\begin{bmatrix} 2.6986 & 1.6548 \\ 1.6548 & 8.9978 \end{bmatrix}$	$\begin{bmatrix} 2.5796 & 1.5845 \\ 1.5845 & 8.8708 \end{bmatrix}$	$\begin{bmatrix} 2.5761 & 1.5820 \\ 1.5820 & 8.8674 \end{bmatrix}$	$\begin{bmatrix} 2.5770 & 1.5823 \\ 1.5823 & 8.8671 \end{bmatrix}$

Note: [a]$E[SB]$ denotes the number of busy servers in the system $\equiv E[\mathrm{FES}] + E[\mathrm{SOS}]$.

Table 3. The optimal value (c^*, μ_1^*, μ_2^*) and the minimum expected cost value for various value of λ, θ and σ, while c^* is obtained at initial value $(\mu_1, \mu_2) = (20, 10)$.

$(\lambda, \theta, \sigma)$	$(5, 0.2, 10)$	$(10, 0.2, 10)$	$(20, 0.2, 10)$	$(5, 0.8, 10)$	$(10, 0.8, 10)$	$(20, 0.8, 10)$
c^*	1	1	2	1	2	4
(μ_1^*, μ_2^*)	[11.8535, 4.26058]	[22.0254, 7.71166]	[22.7810, 7.77980]	[14.7456, 9.53810]	[15.1755, 9.76186]	[16.2154, 10.3483]
$F(c^*, \mu_1, \mu_2)$	628.502	947.158	1200.47	866.965	1129.98	1652.39
E[Orbit]	2.56395	4.79291	3.93227	3.54492	2.88299	1.80663
E[SB]	0.65653	0.71337	1.39208	0.75845	1.47847	2.77955
	(10, 0.2, 15)	(10, 0.5, 15)	(10, 0.8, 15)	(20, 0.2, 15)	(20, 0.5, 15)	(20, 0.8, 15)
c^*	1	2	2	2	3	3
(μ_1^*, μ_2^*)	[21.4641, 7.60174]	[13.8213, 7.19819]	[14.9257, 9.61603]	[22.2164, 7.65119]	[18.2749, 9.41276]	[19.6974, 12.6312]
$F(c^*, \mu_1, \mu_2)$	925.598	1009.21	1118.22	1181.57	1417.74	1561.528
E[Orbit]	4.32420	2.23082	2.62489	3.52100	2.41688	2.93133
$E[SB]$	0.72899	1.41814	1.50193	1.42303	2.15678	2.28207
	(10, 0.2, 5)	(10, 0.2, 10)	(10, 0.2, 15)	(10, 0.8, 5)	(10, 0.8, 10)	(10, 0.8, 15)
c^*	1	1	1	2	2	2
(μ_1^*, μ_2^*)	[23.4453, 8.02222]	[22.0254, 7.71166]	[21.4641, 7.60174]	[15.8259, 10.1461]	[15.1755, 9.76186]	[14.9257, 9.61603]
$F(c^*, \mu_1, \mu_2)$	1003.92	947.158	925.598	1161.21	1129.98	1118.22
E[Orbit]	2.31269	4.79291	4.32420	3.55994	2.88299	2.62489
E[SB]	2.67521	0.71337	0.72899	1.42036	1.47847	1.50193

and the two service rates so that the total expected cost is minimised. Efficient search approaches were presented to obtain the optimal number of channels and the optimal service rates. We performed a sensitivity analysis of the joint optimal values (c^*, μ_1^*, μ_2^*) with respect to specific values of λ, θ and σ. The results would be useful and significant for modelling banking service systems, computer job processing, automatic machine quality control service channels and many related other applications.

Acknowledgements

The authors are grateful to the three referees and associate editor whose constructive comments have led to a substantial improvement in the presentation of this article.

Notes on contributors

Jau-Chuan Ke is a Professor of Applied Statistics at National Taichung Institute of Technology, Taiwan. He received his MS in Applied Mathematics at National Cheng-Kung University, Taiwan, and his PhD in Applied Mathematics at National Chung-Hsing University, Taiwan. His research interests include stochastic processes, queueing theory, reliability and applied statistics.

Chia-Huang Wu received her MS degree in Applied Mathematics from National Chung-Hsing University. Currently, she is a PhD candidate at the Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan, ROC.

Wen Lea Pearn is a Professor of Operations Research and Quality Assurance at National Chiao Tung University, Taiwan. He received his PhD in Operations Research from the University of Maryland, College Park, Maryland, USA. He worked at AT&T Bell Laboratories as a Quality Research Scientist before joining National Chiao Tung University. His research interests include process capability, network optimisation, production management and queueing theory. His publications have appeared in the *Journal of the Royal Statistical Society*, *Journal of Quality Technology*, *Journal of Applied Statistics*, *Statistics*, *Journal of the Operational Research Society*, *European Journal of Operations Research*, *Operations Research Letters*, *Omega*, *Networks*, *International Journal of Production Research*, *International Journal of Systems Science*, *Journal of Computational and Applied Mathematics*, *Quality & Reliability Engineering International*, *IEEE Transactions on Semiconductor Manufacturing*, *IEEE Transactions on Electronics Packaging Manufacturing* and others.

References

Al-Jararha, J., and Madan, K.C. (2003), 'An M/G/1 Queue with Second Optional Service with General Service Time Distribution', *Information and Management Sciences*, 14, 47–56.

Amador, J., and Artalejo, J.R. (2007), 'On the Distribution of the Successful and Blocked Events in the M/M/*c* Retrial Queue: A Computational Approach', *Applied Mathematics and Computation*, 190, 1612–1626.

Artalejo, J.R. (1999a), 'A Classified Bibliography of Research on Retrial Queues: Progress in 1990–1999', *Top*, 7, 187–211.

Artalejo, J.R. (1999b), 'Accessible Bibliography on Retrial Queues', *Mathematical and Computer Modelling*, 30, 1–6.

Artalejo, J.R., Chakravarthy, S.R., and Lopez-Herrero, M.J. (2007a), 'The Busy Period and the Waiting Time Analysis of a MAP/M/c Queue with Finite Retrial Group', *Stochastic Analysis and Applications*, 25, 445–469.

Artalejo, J.R., and Choudhury, G. (2004), 'Steady State Analysis of an M/G/1 Queue with Repeated Attempts and Two-phase Service', *Quality Technology and Quantitative Management*, 1, 189–199.

Artalejo, J.R., Economou, A., and Lopez-Herrero, M.J. (2007b), 'Algorithmic Approximations for the Busy Period Distribution of the M/M/*c* Retrial Queue', *European Journal of Operational Research*, 176, 1687–1702.

Artalejo, J.R., Economou, A., and Lopez-Herrero, M.J. (2007c), 'Algorithmic Analysis of the Maximum Queue Length in a Busy Period for the M/M/*c* Retrial Queue', *INFORMS Journal on Computing*, 19, 121–126.

Artalejo, J.R., and Lopez-Herrero, M.J. (2000), 'On the Busy Period of the M/G/1 Retrial Queue', *Naval Research Logistics*, 47, 115–127.

Artalejo, J.R., and Pozo, M. (2002), 'Numerical Calculation of the Stationary Distribution of the Main Multiserver Retrial Queue', *Annals of Operations Research*, 116, 41–56.

Breuer, L., Dudin, A.N., and Klimenok, V.I. (2002), 'A retrial BMAP/PH/N System', *Queueing Systems*, 40, 433–457.

Bright, L., and Taylor, P.G. (1995), 'Calculating the Equilibrium Distribution in Level Dependent Quasi-Birth-and-death Processes', *Communications in Statistics – Stochastic Models*, 11, 497–525.

Chakravarthy, S.R., and Dudin, A.N. (2002), 'Multiserver Retrial Queue with BMAP Arriving and Group Services', *Queueing Systems*, 42, 5–31.

Chong, E.K.P., and Zak, S.H. (2001), *An Introduction to Optimization* (3rd ed.), Hoboken, NJ: John Wiley and Sons.

Choudhury, G., and Deka, K. (2008), 'An M/G/1 Retrial Queueing System with Two Phases of Service Subject to the Server Breakdown and Repair', *Performance Evaluation*, 65, 714–724.

Choudhury, G., and Madan, K.C. (2005), 'A Two-stage Batch Arrival Queueing System with a Modified Bernoulli Schedule Vacation Under N-policy', *Mathematical and Computer Modelling*, 42, 71–85.

Choudhury, G., and Paul, M. (2006), 'A Batch Arrival Queue with a Second Optional Service Channel Under N-policy', *Stochastic Analysis and Applications*, 24, 1–21.

Choudhury, G., and Tadj, L. (2009), 'An M/G/1 Queue with Two Phases of Service Subject to the Server Breakdown and Delayed Repair', *Applied Mathematical Modelling*, 33, 2699–2709.

Falin, G.I. (1983), 'Calculations of Probability Characteristics of a Multiline System with Repeated Calls', *Moscow University Computational Mathematics and Cybernetics*, 1, 43–49.

Falin, G.I., and Templeton, J.G.C. (1997), *Retrial Queues*, London: Chapman and Hall.

Gomez-Corral, A. (2006), 'A Bibliographical Guide to the Analysis of Retrial Queues Through Matrix Analytic Techniques', *Annals of Operations Research*, 141, 177–207.

Ke, J.-C. (2008), 'An $M^{[x]}$/G/1 System with Startup Server and J Additional Options for Service', *Applied Mathematical Modelling*, 32, 443–458.

Ke, J.-C., and Chang, F.-M. (2009), '$M^{[x]}/(G_1, G_2)/1$ Retrial Queue Under Bernoulli Vacation Schedules with General Repeated Attempts and Starting Failures', *Applied Mathematical Modelling*, 33, 3186–3196.

Krishna, K.B., Rukmani, R., and Thangaraj, V. (2009), 'An M/M/*c* Retrial Queueing System with Bernoulli Vacations', *Journal of Systems Science and Systems Engineering*, 18, 222–242.

Krishna, K.B., and Raja, J. (2006), 'On Multiserver Feedback Retrial Queues with Balking and Control Retrial Rate', *Annal of Operations Research*, 141, 211–232.

Kulkarni, V.G., and Liang, H.M. (1997), 'Retrial Queues Revisited', in *Frontiers in Queueing: Models and Applications in Science and Engineering*, ed. J.H. Dshalalow, Boca Raton: CRC Press, pp. 19–34.

Latouche, G. and Ramaswami, V. (1999), *Introduction to Matrix Analytic Methods in Stochastic Modeling* (ASA-SIAM Series on Statistics and Applied Probability), Philadelphia, PA: Society for Industrial and Applied Mathematics.

Madan, K.C. (2000), 'An M/G/1 Queue with Second Optional Service', *Queueing Systems*, 34, 37–46.

Madan, K.C. (2001), 'On a Single Server Queue with Two-Stage Heterogeneous Service and Deterministic Server Vacations', *International Journal of Systems Science*, 32, 837–844.

Medhi, J. (2002), 'A Single Server Poisson Input Queue with a Second Optional Channel', *Queueing Systems*, 42, 239–242.

Neuts, M.F. (1981), *Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach*, Baltimore, USA: The John Hopkins University Press.

Neuts, M.F., and Rao, B.M. (1990), 'Numerical Investigation of a Multiserver Retrial Model', *Queueing Systems: Theory and Applications*, 7, 169–190.

Stepanov, S.N. (1999), 'Markov Models with Retrials: The Calculation of Stationary Performance Measures Based on the Concept of Truncation', *Mathematical and Computer Modelling*, 30, 207–228.

Tadj, L., Choudhury, G., and Tadj, C. (2006a), 'A Quorum Queueing System with a Random Setup Time Under N-policy and with Bernoulli Vacation Schedule', *Quality Technology and Quantitative Management*, 3, 145–160.

Tadj, L., Choudhury, G., and Tadj, C. (2006b), 'A Bulk Quorum Queueing System with a Random Setup Time Under N-policy and with Bernoulli Vacation Schedule', *Stochastics: An International Journal of Probability and Stochastics Processes*, 78, 1–11.

Wang, J. (2004), 'An M/G/1 Queue with Second Optional Service and Server Breakdowns', *Computers and Mathematics with Applications*, 47, 1713–1723.

Yang, T., and Templeton, J.G.C. (1987), 'A Survey on Retrial Queues', *Queueing Systems: Theory and Applications*, 2, 201–233.