Journal of Computational and Applied Mathematics 235 (2010) 547–562

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Computational algorithm and parameter optimization for a multi-server system with unreliable servers and impatient customers

Chia-Huang Wu [a], Jau-Chuan Ke [b,*]

[a] Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan, ROC

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

ARTICLE INFO

Article history:
Received 20 October 2008
Received in revised form 27 April 2010

Keywords:
Cost
Balk
Quasi-Newton method
Renege
Matrix analytic approach

ABSTRACT

We consider an infinite capacity M/M/c queueing system with c unreliable servers, in which the customers may balk (do not enter) and renege (leave the queue after entering). The system is analyzed as a quasi-birth-and-death (QBD) process and the necessary and sufficient condition of system equilibrium is obtained. System performance measures are explicitly derived in terms of computable forms. The useful formulae for computing the rate matrix and stationary probabilities are derived by means of a matrix analytical approach. A cost model is derived to determine the optimal values of the number of servers, service rate and repair rate simultaneously at the minimal total expected cost per unit time. The parameter optimization is illustrated numerically by the Quasi-Newton method.

1. Introduction

In this paper, the quasi-birth-and-death (QBD) process and the matrix analytic method are used to analyze an infinite capacity multi-server M/M/c queue with unreliable servers, balking and reneging customers. The computational algorithm of stationary probability vectors and optimization of parameters are developed.

In most studies on queueing systems, the customers always wait in the system until his service is completed. In many practical systems, such as telephone switchboard customers, hospital emergency rooms' handling of critical patients, and perishable goods storage inventory systems, the customers may become impatient and leave (i.e., balk or renege) the system without getting services when the waiting time is intolerable. For example, for a call-in customer who cannot get service immediately by the server, he/she is told how long he/she needs to wait. The customer might hang up (balk) or hold on (non-balk and waiting). This is a balking behavior of the customer when the queue length or waiting time is too long. In addition, a waiting customer might hang up (renege) if he/she becomes impatient.

Someone who wants to buy a train ticket (or meal ticket) might decide not entering the system (balk) if the waiting line is too long. On the other hand, as a customer waiting in the queue, he/she might leave the queue (renege) and choose an automat (or instant food). Interesting examples of the occurrence of balking and reneging in air defence systems can be found in [1,2]. In addition, balking and reneging are also common phenomena in telecommunication networks and machine repair problems (see [3,4]).

Queueing models with balking, or reneging, or both have attracted much attention from numerous researchers since Haight [5,6]. The extensions of their basic model can be found in [7–11,3,12]. Later, Wang and Chang [13] examined

* Corresponding author. Tel.: +886 4 22196077; fax: +886 4 22196331.
E-mail address: jauchuan@ntit.edu.tw (J.-C. Ke).

0377-0427/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2010.06.005

a finite capacity M/M/R queueing system with balking, reneging, and server breakdowns. Al-Seedy [14] proposed a transient solution of the non-truncated queue M/M/2 involving balking, and an additional server for longer queues. The maximum likelihood estimates and confidence intervals of an M/M/R/N queue with balking and heterogeneous servers were investigated in [15]. Recently, Ke [16] gave the operating characteristics analysis of the $M^{[x]}/G/1$ system with a variant vacation policy and balking using supplementary variable technique. Yue and Yue [17] analyzed a finite capacity M/M/c/N queueing system with balking, reneging and synchronous vacations. In [17] they derived the steady-state probability vector expressed as the inverse of two matrices with the blocked matrix forms.

Alternatively, many queueing systems were studied with assumption that the server would never fail. In practice, we often encounter cases where the server may fail and can be repaired. Recently, several researches devoted efforts to investigate the impact of unreliable servers (server breakdowns), in which the server subject to unpredictable breakdowns and can be repaired. Past work of unreliable-server queueing models may be divided into two categories, according to whether the system was studied from finite capacity or infinite capacity. Plenty studies in the first category focused on machine repair problems, interested readers may refer to Ke and Wang [3], references there in. Recently, Wang et al. [4] studied the models in [3] and then optimized the number of servers, balking and reneging rates using direct search and steep decent methods. The second category of authors dealing with papers treating the problem of infinite capacity. A pioneering work in this queueing situation is investigated in [18], who first introduced the concept of service interruptions (server breakdowns). Gaver [18] studied an M/G/1 queue with service interruptions by using the embedded Markovian chain. Tang [19] studied some queueing problems of the system and some reliability problems of the broken server for a single unreliable-server M/G/1 queueing system. The reliability measures, were examined in [20], for the ordinary M/G/1 queue with channel breakdowns and vacations. Wang [21] investigated the reliability behavior of the ordinary M/G/1 queue with server breakdowns and second optional service. Wang et al. [22] used maximum entropy principle to approximate the steady-state probability distributions of $M^{[x]}/M/1$ with server breakdowns and vacations. Recently, Choudhury and Tadj [23] generalized this type of model by introducing the concept of a server breakdown and a delay-repair-period. An M/G/1 retrial queueing system with two phases of service subject to the server breakdown and repair was investigated in [24], who derived the queue size distribution at a random epoch and departure epoch using supplementary variable technique, various system performance measures are also presented. As related works with control policy, the readers can refer to the excellent survey in [25]. Recently, Ke [26] derived the system characteristics and examined the optimal NT policies for M/G/1 system with server breakdown and startup using stochastic decomposition property. Various system performance measures and sensitivity investigations based on the optimal threshold N at a minimum cost, were studied in [27], for the M/G/1 queueing system under N policy with server startup and breakdowns.

Existing unreliable-server queues, including those above, mainly focused on *finite capacity multi-server system* or *infinite capacity single-server system*. Because more complicated structure of the stochastic processes required describes system behaviors, the infinite capacity multi-server queue with unreliable servers is known to be analytically intractable. Analytic steady-state solutions of an infinite capacity M/M/c queue with unreliable servers have not been found. This motivates us to investigate an infinite capacity multi-server queue M/M/c type with unreliable servers, balking and reneging customers.

In this paper, we consider an infinite capacity M/M/c queueing system with unreliable servers and impatient customers. Customers arrive according to a Poisson process with parameter λ and their service times are provided by c unreliable servers, in which the service times are assumed to be exponentially distributed with mean $1/\mu$. A customer on arrival finds n customers and i breakdown (unavailable) servers in the system, either decides to enter the queue with probability $b_{i,n}$ or balks (do not enter) with probability $1 - b_{i,n}$. If the service is unoccupied or is not interrupted by a breakdown, an arriving customer immediately starts getting the service. But if a customer enters the queue, it may get impatient and leave the queue without getting service. After entering the queue, each customer will wait a certain time t for service to begin, may leave it without being served. This time t is a random variable which is assumed to be distributed according to the exponential distribution with parameter r. We assume that customers arrive at the server form a single waiting line and are served in the order of their arrivals; that is, the first-come, first served discipline. Whenever the server is working, it is assumed that the server can break down at any time with a Poisson breakdown rate α. Whenever the server fails, it is immediately repaired at a repair rate β, where the repair times are assumed to be exponentially distributed. The server can break down even if no customers are in the system. Each server can serve only one customer at a time, and that the service is independent of the arrival of the customers. Service is additive and allowed to be interrupted if the server breaks down, and the server is immediately repaired. Once the broken down server is repaired, it immediately returns to serve the customer. Although no service occurs during the repair period of all broken down servers, customers continue to arrive according to a Poisson process.

The paper is organized as follows; In Section 2, mathematical model and the quasi-birth-and-death (QBD) model of an infinite capacity M/M/c queue with unreliable servers and impatient customers are set up. The matrix-geometric property (matrix analytic method) is used to calculate rate matrix in Section 3. Section 4 we derive an efficient algorithm to stationary probabilities by matrix-geometric method. Some system performance measures are derived in Section 5. In Section 6, a cost model is developed to determine the optimal values of number of channels, service rate and repair rate, simultaneously, in order to minimize the total expected cost per unit time. We use direct search method and Quasi-Newton method to implement the optimization tasks. Some numerical examples are provided to illustrate the two optimization methods. Section 7 concludes.

2. Mathematical model

We consider an infinite capacity M/M/c queueing system with balking, reneging and server breakdowns. The state of the system can be described by the pair (i, n), $i = 0, 1, 2, \ldots, c$, $n = 0, 1, 2, \ldots$, where i denotes the number of broken down servers, and n denotes the number of customers in the system.

According to system assumptions, a customer who on arrival finds n customers in the system, either decides to enter the queue with probability $b_{i,n}$ or balk with probability $1 - b_{i,n}$, where $b_{i,n}$ are defined as follows

$$b_{i,n} = \begin{cases} 1, & 0 \le n \le c - i - 1 \\ \rho(1-\rho)^{n-(c-i)+1}, & 0 \le c - i \le n. \end{cases}$$

If n is less than or equal to $(c - i)$, the phenomenon of reneging will not occur and the customers which upon the server will get services immediately. Otherwise, there are $[n - (c - i)]$ customers have to wait in the queue. Furthermore, since the waiting time is assumed to be distributed according to an exponential distribution with mean $1/\gamma$, the average reneging rate in this state is given by $\gamma_n = [n - (c - i)]\gamma$.

In steady-state, we define

$P_i(n) \equiv$ probability that there are n customers in the system when there are i servers broken down, where $i = 0, 1, 2, \ldots, c$ and $n = 0, 1, 2, \ldots$

and $\Pi = [\Pi_0, \Pi_1, \ldots, \Pi_n, \Pi_{n+1}, \ldots]$ with $\Pi_n = [P_0(n), P_1(n), \ldots, P_c(n)]$, $n = 0, 1, 2, \ldots$ denotes the steady-state probability vector.

Referring to the state-transition-rate diagram as shown in Fig. 1, it is noted that changes in the state of the system may occur due to one of the following events: (1) a non-balking arrival, (2) a customer reneges, (3) a server breakdown, (4) a repair completion, or (5) a service completion. The first two events depend upon the number of customers n in the system (also see the preceding definition of $b_{i,n}$ and γ_n). As n increases, the non-balking rate $b_{i,n}$ decreases progressively. When n is sufficiently large, most new arriving customers would balk and do not enter the system. That is, the likelihood of a non-balking (successful) arrival is small and converges to a very small value when $n \ge N$, where N is a sufficiently large. Therefore, the balking rate is reasonably modifies as

$$b_{i,n} = \begin{cases} 1, & 0 \le n \le c - i - 1 \\ \rho(1-\rho)^{n-(c-i)+1}, & c - i \le n \le N \\ \rho(1-\rho)^{N-(c-i)+1}, & n > N. \end{cases}$$

Based on aforementioned observations, the balking rate is larger when $n \ge N$, the increasing of the number of customers in the system is limited and consequently the reneging rate is restricted as well. For analysis, the reneging rate $\gamma_{i,n}$ is redefined as

$$\gamma_n = \begin{cases} 0, & n < c - i \\ \min\{[n - (c - i)], [N - (c - i)]\}\gamma, & n \ge c - i. \end{cases}$$

From [28], the infinitesimal generator Q of the QBD describing the M/M/c queueing system with balking, reneging and c non-reliable servers as

$$Q = \begin{bmatrix} A_0 & B_0 & & & & & & & \\ C_1 & A_1 & B_1 & & & & & & \\ & C_2 & A_2 & B_2 & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & C_{N-1} & A_{N-1} & B_{N-1} & & & \\ & & & & C_N & A_N & B_N & & \\ & & & & & C_N & A_N & B_N & \\ & & & & & & \ddots & \ddots & \ddots \end{bmatrix}. \quad (1)$$

The sub-matrices B_k $(0 \le k \le N)$, A_k $(0 \le k \le N)$, and C_k $(1 \le k \le N)$ are diagonal matrices of order $(c + 1)$ which defined by

$$B_k = \begin{cases} \mathrm{diag}(\underbrace{\lambda, \lambda, \ldots, \lambda}_{\#=c-k}, \underbrace{\lambda b_{c-k,k}, \lambda b_{c-k+1,k}, \ldots, \lambda b_{c,k}}_{\#=k+1}), & 0 \le k \le c \\ \mathrm{diag}(\lambda b_{0,k}, \lambda b_{1,k}, \ldots, \lambda b_{c,k}), & c + 1 \le k \le N \end{cases}$$

$$C_k = \begin{cases} \mathrm{diag}(\underbrace{k\mu, k\mu, \ldots, k\mu}_{\#=c-k}, \underbrace{k\mu, (k-1)\mu + \gamma, \ldots, k\gamma}_{\#=k+1}), & 0 \le k \le c \\ \mathrm{diag}(c\mu + (k-c)\gamma, (c-1)\mu + (k-c+1)\gamma, \ldots, k\gamma), & c + 1 \le k \le N \end{cases}$$

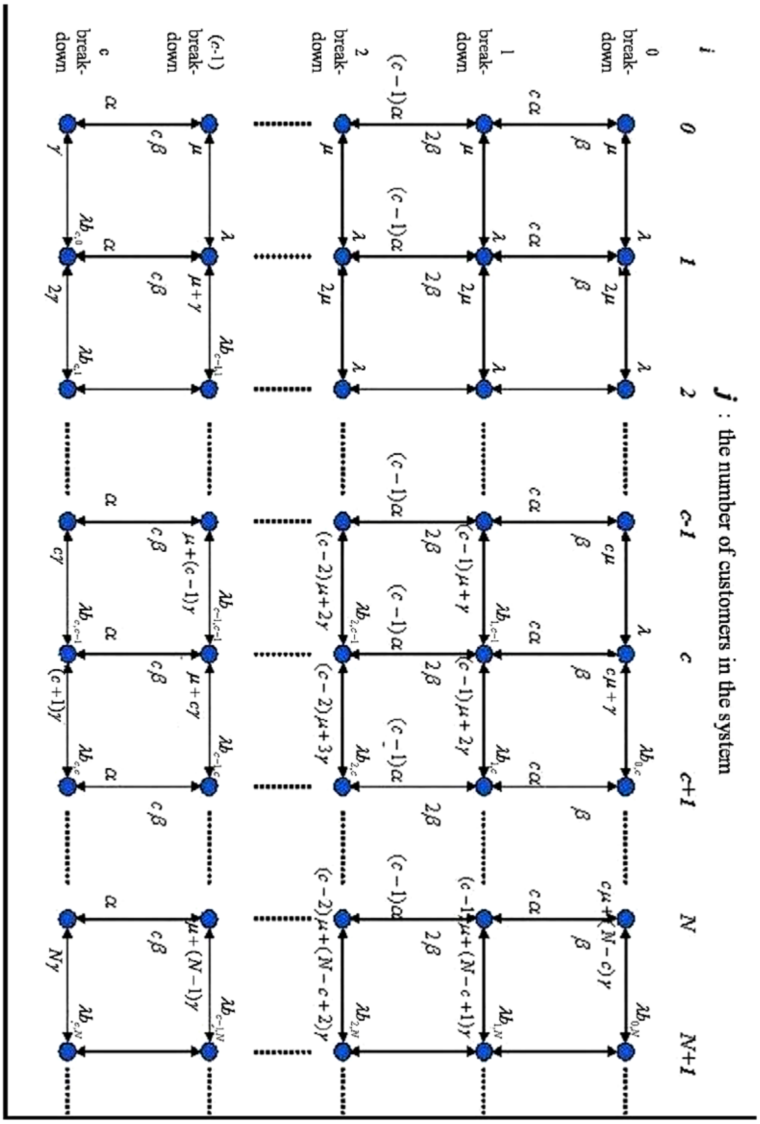

Fig. 1. State-transition-rate diagram for an infinite capacity M/M/c queueing system with unreliable servers and impatient customers.

where "diag(V)" denotes a diagonal matrix with diagonal elements equal to V.

$$A_k = \begin{bmatrix} a_{k,0} & c\alpha & & & & \\ \beta & a_{k,0} & (c-1)\alpha & & & \\ & 2\beta & a_{k,2} & (c-2)\alpha & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & (c-1)\beta & a_{k,c-1} & \alpha \\ & & & & c\beta & a_{k,c} \end{bmatrix}, \quad k = 0, 1, 2, \ldots, N$$

with diagonal elements

$$a_{k,s} = \begin{cases} -[\lambda + (c-s)\alpha + s\beta + k\mu], & 0 \le s \le c-k-1 \\ -[\lambda b_{s,k} + (c-s)\alpha + s\beta + (c-s)\mu + (s+k-c)\gamma], & c-k \le s \le c \end{cases} \quad \text{if } 0 \le k \le c$$

$$a_{k,s} = -[\lambda b_{s,k} + (c-s)\alpha + s\beta + (c-s)\mu + (s+k-c)\gamma], \quad 0 \le s \le c \text{ if } c+1 \le k \le N.$$

The steady-state probability vector Π be the unique solution to $\Pi Q = \mathbf{0}$ and $\sum_{n=0}^{\infty} \Pi_n e = 1$, where e is a column vector with dimension $(c+1)$ and all elements equal to 1. We note that the vector $\Pi = [\Pi_0, \Pi_1, \ldots, \Pi_N, \Pi_{N+1}, \ldots]$ with the following properties

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

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

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

From [28,29], we know that R is given by $\lim_{n\to\infty} R_n$, where the sequence $\{R_n\}$ is defined by

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

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

Algorithm: Quasi Progression Algorithm

INPUT B_N, A_N, C_N, e is a $(c+1)$ column vector with all elements equal to one, and tolerance ε
OUTPUT approximate solution R
Step 1 Set $R = \mathbf{0}$
Step 2 while $|e^T(B_N + RA_N + R^2C_N)e| \geq \varepsilon$ do step 3–4
Step 3 Set $R_{\text{new}} = -B_N A_N^{-1} - R^2 C_N A_N^{-1}$
Step 4 Set $R = R_{\text{new}}$
Step 5 OUTPUT

It is also well known (Theorem 3.1.1 of [28]) that the steady-state probability vector exists if and only if

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

where x is the invariant probability of the matrix $F = C_N + A_N + B_N$. x satisfies $xF = \mathbf{0}$ and $xe = 1$. First we solve $xF = \mathbf{0}$, where $x = [x_0, x_1, \ldots, x_c]$. We can write following $(c+1)$ equations

$$-cx_0\alpha + x_1\beta = 0, \tag{6a}$$

$$(c-i+1)x_{i-1}\alpha - x_i[(c-i)\alpha + i\beta] + (i+1)x_{i+1}\beta = 0, \quad 1 \leq i \leq c-1, \tag{6b}$$

$$x_{c-1}\alpha - cx_c\beta = 0. \tag{6c}$$

Eq. (6a) implies that $x_1 = \frac{c\alpha}{\beta}x_0$, and solving (6b)–(6c) recursively, we get

$$x_{i+1} = \frac{(c-i)\alpha}{(i+1)\beta}x_i, \quad i = 1, \ldots, c-1. \tag{7}$$

Finally, we have

$$x_{i+1} = \frac{(c-i)\alpha}{(i+1)\beta}x_i = \frac{(c-i)(c-i-1)}{(i+1)i}\left(\frac{\alpha}{\beta}\right)^2 x_{i-1} = \cdots = \binom{c}{i+1}\left(\frac{\alpha}{\beta}\right)^{i+1} x_0, \quad i = 1, \ldots, c-1. \tag{8}$$

Also using the normalization condition $x_0 + x_1 + \cdots + x_{c-1} + x_c = 1$, we can determine x_0 as

$$x_0 = \left[\sum_{i=0}^{c}\binom{c}{i}\left(\frac{\alpha}{\beta}\right)^i\right]^{-1} = \left(1 + \frac{\alpha}{\beta}\right)^{-c}. \tag{9}$$

Substituting B_N and C_N into Eq. (5) and doing some routine manipulations, then we have

$$\lambda\sum_{i=0}^{c}\binom{c}{i}\left(\frac{\alpha}{\beta}\right)^i b_{i,N} < \sum_{i=0}^{c}\{(c-i)\mu + [N-(c-i)]\gamma\}\binom{c}{i}\left(\frac{\alpha}{\beta}\right)^i, \tag{10}$$

or equivalent

$$\lambda E_i[b_{i,N}] < E[H]\mu + (N - E[H])\gamma, \tag{11}$$

where

$$E_i[b_{i,N}] = \sum_{i=0}^{c}\binom{c}{i}\left(\frac{\alpha}{\beta}\right)^i b_{i,N}\left(1+\frac{\alpha}{\beta}\right)^{-c}, \qquad E[H] = \sum_{i=0}^{c}(c-i)\binom{c}{i}\left(\frac{\alpha}{\beta}\right)^i\left(1+\frac{\alpha}{\beta}\right)^{-c}$$

mean the average of entering probability (no balking) and the average number of normal (no breakdown) servers, respectively. As $\gamma = 0$ and $b_{i,N} = 1$ for all i, Eq. (11) can be reduced to stability condition of the ordinary M/M/c queueing system with unreliable servers (i.e., $\lambda < E[H]\mu$).

3. Algorithm solution

Under the stability condition, by solving the equation $PQ = \mathbf{0}$ with the normalization condition, we obtain

$$\Pi_0 A_0 + \Pi_1 C_1 = \mathbf{0}, \tag{12a}$$

$$\Pi_{n-1} B_{n-1} + \Pi_n A_n + \Pi_{n+1} C_{n+1} = \mathbf{0}, \quad 1 \le n \le N-1, \tag{12b}$$

$$\Pi_{N-1} B_{N-1} + \Pi_N A_N + \Pi_N R C_N = \mathbf{0}, \tag{12c}$$

$$\Pi_N R^{n-N-1} B_N + \Pi_N R^{n-N} A_N + \Pi_N R^{n-N+1} C_N = \mathbf{0}, \quad N+1 \le n, \tag{12d}$$

$$\sum_{n=0}^{\infty} \Pi_n e = 1. \tag{13}$$

After doing routine substitutions to (12a)–(12c), we have

$$\begin{aligned} \Pi_0 &= \Pi_1 C_1 (-A_0)^{-1} = \Pi_1 \phi_1, \\ \Pi_{n-1} &= \Pi_n C_n [-(\phi_{n-1} B_{n-2} + A_{n-1})]^{-1} = \Pi_n \phi_n, \quad 2 \le n \le N, \end{aligned} \tag{14}$$

and

$$\Pi_N \phi_N B_{N-1} + \Pi_N A_N + \Pi_N R C_N = \mathbf{0}. \tag{15}$$

Consequently, Π_n $(0 \le n \le N-1)$ in Eq. (14) can be written in terms of Π_N as $\Pi_n = \Pi_N \Pi_{i=N}^{n+1} \phi_i$, $n = 0, 1, 2, \ldots, N-1$ and the rest steady-state vectors $\Pi_N, \Pi_{N+1}, \ldots$ can be calculated recursively as $\Pi_n = \Pi_N R^{n-N}$, for $n \ge N$. Once Π_N is determined, the steady-state solutions $\Pi = [\Pi_0, \Pi_1, \ldots, \Pi_N, \Pi_{N+1}, \ldots]$ are obtained. The vector Π_N is given by solving Eq. (15) with the following normalization condition.

$$\begin{aligned} \sum_{n=0}^{\infty} \Pi_n e &= [\Pi_0 + \Pi_1 + \cdots + \Pi_{N-1} + \Pi_N + \Pi_{N+1} + \Pi_{N+2} + \cdots] e \\ &= [\Pi_N \Pi_{i=N}^{1} \phi_i + \Pi_N \Pi_{i=N}^{2} \phi_i + \cdots + \Pi_N \Pi_{i=N}^{N} \phi_i + \Pi_N + \Pi_N R + \Pi_N R^2 + \cdots] e \\ &= \Pi_N \left[\sum_{n=1}^{N} \Pi_{i=N}^{n} \phi_i + (I - R)^{-1} \right] e = 1. \end{aligned} \tag{16}$$

Solving Eqs. (15) and (16) 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 (14) and $[\Pi_{N+1}, \Pi_{N+2}, \Pi_{N+3}, \ldots]$ are gained by the formula $\Pi_n = \Pi_N R^{n-N}$, $n \ge N+1$. The solution procedure of the steady-state probabilities is summarized as below:

Algorithm: Recursive Solver

INPUT $c, N, B_0, B_1, \ldots, B_N, A_0, A_1, \ldots, A_N, C_1, C_2, \ldots, C_N, R$.
OUTPUT approximate solution $\Pi_0, \Pi_1, \Pi_2, \ldots$
Step 1 set $\phi_1 = C_1(-A_0)^{-1}$
Step 2 for $i = 2$ to N
  Step 3 set $\phi_i = C_i[-(\phi_{i-1} B_{i-2} + A_{i-1})]^{-1}$
Step 4 end
Step 5 for $k = 1$ to N
  Step 6 set $\Phi_k = \Pi_{i=N}^{k} \phi_i$
Step 7 end
Step 8 Solving $\Pi_N \phi_N B + \Pi_N A_N + \Pi_N R C_N = \mathbf{0}$, and $\Pi_N [\sum_{k=1}^{N} \Phi_k + (I - R)^{-1}] e = 1$
Step 9 for $i = 0$ to $N-1$
  Step 10 set $\Pi_i = \Pi_N \Phi_{i+1}$
Step 11 end
Step 12 for $i = N+1$ to infinity
  Step 13 set $\Pi_{i+1} = \Pi_i R$
Step 14 end
Step 15 OUTPUT

4. System performance measures

In this section, we derive some system performance measures of the system such as the expected number of customers in the system (denoted by L_s), the expected number of customers in the queue (denoted by L_q), the expected number of busy, idle and breakdown servers (denoted by $E[B]$, $E[I]$ and $E[D]$, respectively). The expressions for these system performances are given by

$$\begin{aligned} L_s &= \sum_{n=1}^{\infty} n\Pi_n e = \sum_{n=1}^{N-1} n\Pi_n e + N\Pi_N e + (N+1)\Pi_N Re + \cdots \\ &= \sum_{n=1}^{N-1} n\Pi_N \Phi_{n+1} e + N\Pi_N (I-R)^{-1} e + \Pi_N R(I-R)^{-2} e \\ &= \Pi_N \left[\sum_{n=1}^{N-1} n\Phi_{n+1} + N(I-R)^{-1} + R(I-R)^{-2} \right] e \end{aligned} \tag{17}$$

$$\begin{aligned} L_q &= \Pi_1 \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} + \Pi_2 \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 2 \end{bmatrix} + \cdots + \Pi_N \begin{bmatrix} N-c \\ \vdots \\ N-1 \\ N \end{bmatrix} + \Pi_N R \begin{bmatrix} N-c+1 \\ \vdots \\ N \\ N+1 \end{bmatrix} + \cdots \\ &= \sum_{n=1}^{N-1} \Pi_N \Phi_{n+1} u_n + \Pi_N (I-R)^{-1} u_N + \Pi_N R(I-R)^{-1} e \\ &= \Pi_N \left[\sum_{n=1}^{N-1} \Phi_{n+1} u_n + (I-R)^{-1} u_N + R(I-R)^{-1} e \right] \end{aligned} \tag{18}$$

$$E[D] = \sum_{j=0}^{\infty} \Pi_j \begin{bmatrix} 0 \\ 1 \\ \vdots \\ c \end{bmatrix} = \Pi_N \left[\sum_{j=1}^{N} \Phi_j + (I-R)^{-1} \right] \begin{bmatrix} 0 \\ 1 \\ \vdots \\ c \end{bmatrix} \tag{19}$$

$$\begin{aligned} E[I] &= \Pi_0 \begin{bmatrix} c \\ c-1 \\ \vdots \\ 0 \end{bmatrix} + \Pi_1 \begin{bmatrix} c-1 \\ c-2 \\ \vdots \\ 0 \end{bmatrix} + \Pi_2 \begin{bmatrix} c-2 \\ c-3 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \Pi_{c-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= \Pi_N \Phi_1 v_0 + \Pi_N \Phi_2 v_1 + \cdots + \Pi_N \Phi_c v_{c-1} \\ &= \Pi_N \sum_{j=1}^{c} \Phi_j v_{j-1} \end{aligned} \tag{20}$$

$$E[B] = c - E[D] - E[I] \tag{21}$$

where u_n is a $(c+1)$ dimensional column vector with the kth element equal to $\max\{0, n-(c-k+1)\}$ and $v_j = [\underbrace{c-j, c-j-1, \ldots, 1, 0}_{\#=c-j+1}, \underbrace{0, \ldots, 0}_{\#=j}]^{\mathrm{T}}$.

We also give the steady-state availability and using the concept of Ancker and Gafarian [1,2], the expected balking rate, the expected reneging rate and the expected rate of customer loss are obtained as follows:

1. The steady-state availability

$$AV = \Pi_N \left[\sum_{i=1}^{N} \Phi_i + (I-R)^{-1} \right] e_{c+1} \tag{22}$$

where e_{c+1} is a column vector with dimension $(c+1)$ and all elements equal to 1 except that the $(c+1)$th element equal to zero.

2. The expected balking rate

$$\begin{aligned} \text{B.R.} &= \lambda \Pi_0 \boldsymbol{\rho}_0 + \lambda \Pi_1 \boldsymbol{\rho}_1 + \cdots + \lambda \Pi_N \boldsymbol{\rho}_N + \lambda \Pi_N R \boldsymbol{\rho}_N + \lambda \Pi_N R^2 \boldsymbol{\rho}_N \cdots \\ &= \lambda \Pi_N [\Phi_1 \boldsymbol{\rho}_0 + \Phi_2 \boldsymbol{\rho}_1 + \cdots + \Phi_N \boldsymbol{\rho}_{N-1} + (I-R)^{-1} \boldsymbol{\rho}_N] \end{aligned} \tag{23}$$

where $\boldsymbol{\rho}_j$ is defined by $\boldsymbol{\rho}_j = \begin{cases} [\underbrace{0, \ldots, 0, 0}_{\#=c-j}, \underbrace{1-b_{c-j,j}, \ldots, 1-b_{c,j}}_{\#=j+1}]^{\mathrm{T}}, & j \le c \\ [1-b_{0,j}, \ldots, 1-b_{c,j}]^{\mathrm{T}}, & c+1 \le j \le N+1. \end{cases}$

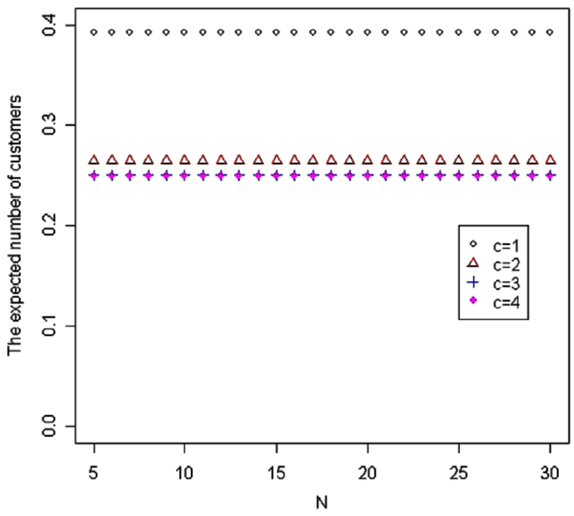

Fig. 2. The effect of N on the expected number of customers in the system.

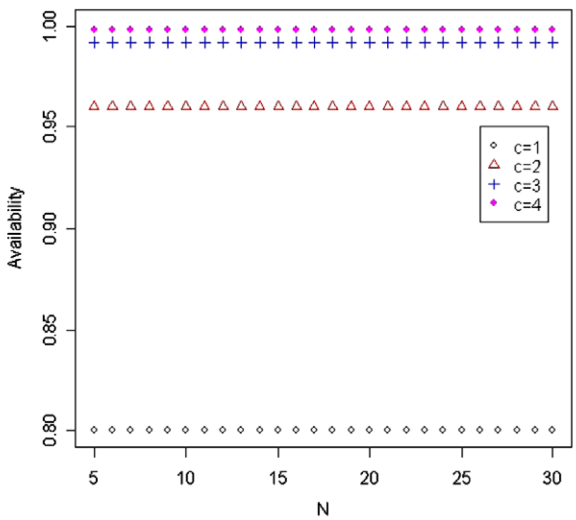

Fig. 3. The effect of N on the steady-state availability.

3. The expected reneging rate

$$\text{R.R.} = \Pi_1\Gamma_1 + \Pi_2\Gamma_2 + \cdots + \Pi_{N-1}\Gamma_{N-1} + \Pi_N\Gamma_N + \Pi_N R\Gamma_N + \Pi_N R^2\Gamma_N \cdots$$
$$= \Pi_N[\Phi_2\Gamma_1 + \Phi_3\Gamma_2 + \cdots + \Phi_N\Gamma_{N-1} + (I-R)^{-1}\Gamma_N] \tag{24}$$

where Γ_j is a column vector with dimension $(c+1)$ and the kth element equal to $\max\{0, j-(c-k+1)\}\gamma$.

4. The expected loss rate L.R. = B.R. + R.R.

To understand how system performance measures listed above vary with N, we now perform some numerical investigation to the measures based on changing the value of N. It should be noted that N initiates with 5 since the number of server which we considered is from 1 to 4 (the readers can refer to system assumptions). As $\alpha = 0.05$, $\beta = 0.2$, $\gamma = 0.5$, $\lambda = 5$ and $\mu = 20$, the numerical illustrations for N versus L_s, AV, L.R. are graphically presented in Figs. 2–4, respectively.

From Figs. 2–4, we observe that the reneging or balking does not result in a commensurate improvement in the system performance measures. Intuitively, N rarely affects the measures. For computation convenience, we adopt $N = 30$ in following numerical examples.

For an infinite capacity M/M/c queueing system with unreliable servers and impatient customers, the numerical results of L_s are obtained by considering the following two cases with different values of c.

Case 1. $\mu = 20$, $\alpha = 0.05$, $\beta = 0.2$, $\gamma = 0.5$, vary λ from 1.0 to 5.0.

Case 2. $\lambda = 2.0$, $\alpha = 0.05$, $\beta = 0.2$, $\gamma = 0.5$, vary μ from 20 to 25.

Results of L_s are depicted in Figs. 5 and 6 for Cases 1–2, respectively.

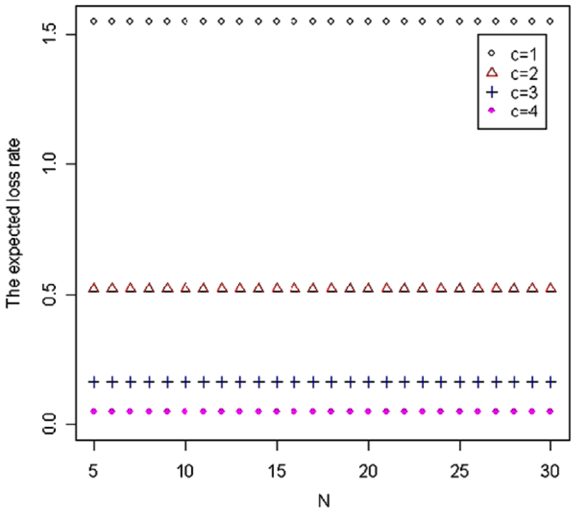

Fig. 4. The effect of N on the expected loss rate.

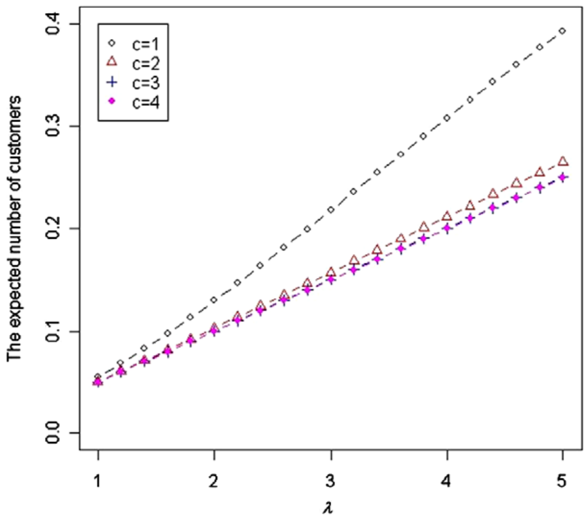

Fig. 5. The effect of λ on the expected number of customers in the system.

One sees from Fig. 5 that L_s drastically increases as λ increases. It reveals from Fig. 6 that L_s drastically decreases as μ increases. To further understand how system performance measures listed above vary with system parameters, we also perform some numerical investigations to the measures based on changing the value of λ, μ, α, and β. The following four cases are performed by considering the different values of c.

Case 3. Availability versus α from 0.05 to 0.45 when $\lambda = 2.5$, $\mu = 3.0$, $\gamma = 0.5$, and $\beta = 0.5$.

Case 4. Availability versus β from 0.1 to 0.5 when $\lambda = 2.5$, $\mu = 3.0$, $\gamma = 0.5$, and $\alpha = 0.05$.

Case 5. The expected loss rate versus λ from 2.5 to 4.5 when $\mu = 5.0$, $\gamma = 0.5$, $\alpha = 0.05$, and $\beta = 0.2$.

Case 6. The expected loss rate versus μ from 3.0 to 5.0 when $\lambda = 2.5$, $\gamma = 0.5$, $\alpha = 0.05$, and $\beta = 0.2$.

The numerical illustrations of the steady-state availability and the expected loss rate are graphically presented in Figs. 7–10 for Cases 3–6, respectively. We observe from Figs. 7 and 8 that the steady-state availability drastically decreases for as α increases or β decreases for $c = 1$, while it is not very insensitive to α or β for $c \geq 2$. From Figs. 9 and 10, it seems that the expected loss rate is more sensitive to the change of λ than the change of μ.

5. Optimization analysis

In this section, we construct the total expected cost function per unit time based on the system performance measures for such a system in which the number of servers (c) is a discrete decision variable, and then the service rate (μ) and the repair

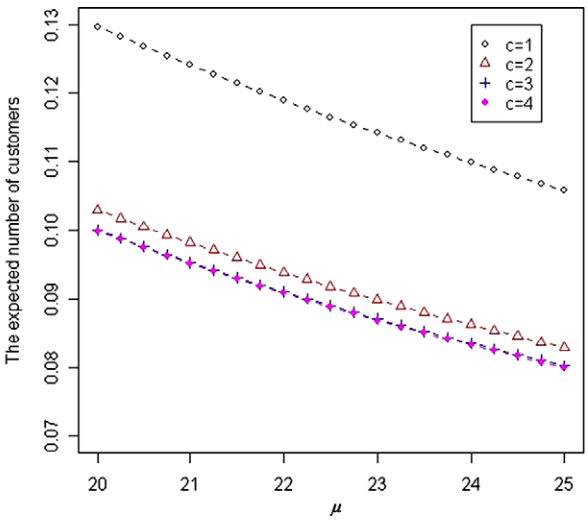

Fig. 6. The effect of μ on the expected number of customers in the system.

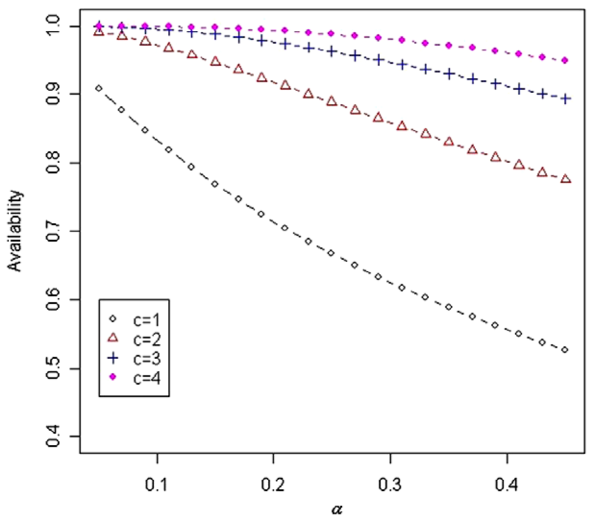

Fig. 7. The effect of α on the steady-state availability.

rate (β) are continuous decision variables. Our main objective is to find the optimum number of servers c^*, and the optimum values of service rate and repair rate (μ^*, β^*) to minimum the cost function. Let us define the following cost elements:

$C_h \equiv$ holding cost per unit time per customer present in the system;
$C_s \equiv$ cost per unit time of providing an service rate μ;
$C_d \equiv$ cost per unit time when one server is broken down;
$C_r \equiv$ cost per unit time of providing a repair rate β;
$C_l \equiv$ loss cost per unit time when one customer balks or reneges;
$C_p \equiv$ fixed cost for purchasing one server.

Using the definition of the cost parameters listed above, the total expected cost function per unit time is given by:

$$F(c, \mu, \beta) = C_h L_s + C_s \mu + C_d E[D] + C_r \beta + C_l \text{L.R.} + C_p c \tag{25}$$

where L_s, $E[D]$, and L.R. are defined previously.

The analytic study of the optimization behavior of the expected cost function would have been an arduous task to undertake since the decision variables appear in an expression which is a highly nonlinear and complex. Based on the preceding formulation, we use a direct search method to compute/find the optimal value of the number of servers, say

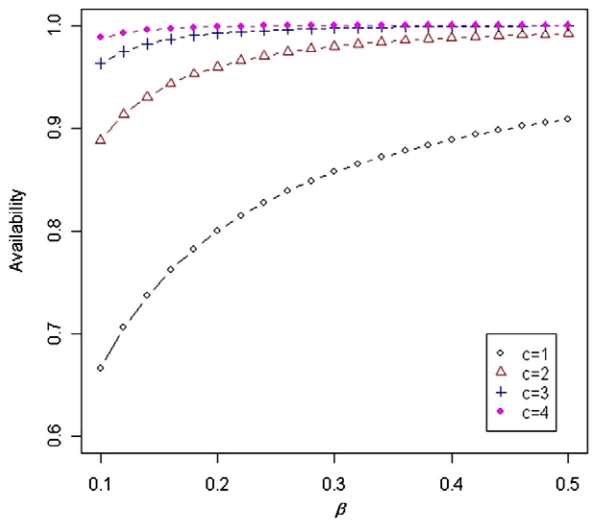

Fig. 8. The effect of β on the steady-state availability.

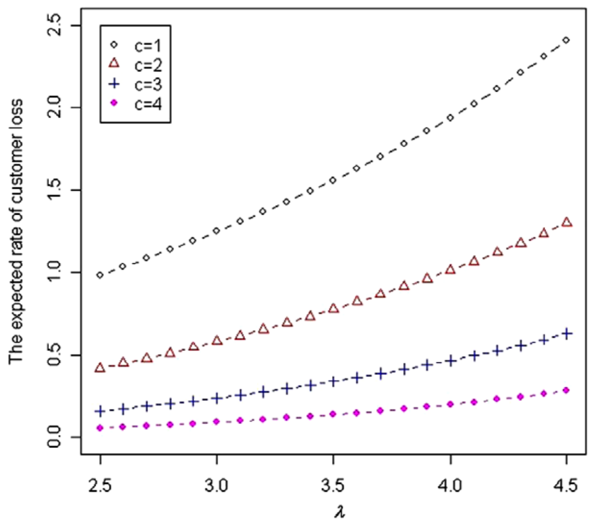

Fig. 9. The effect of λ on the expected loss rate.

c^*, when μ and β are fixed. We then fix c^* and use the Quasi-Newton method to search/adjust the optimal value of μ and β, say μ^* and β^*.

5.1. Direct search method

The optimum value c^* can be determined by the following inequality with satisfying Eq. (10) (stability condition) satisfied

$$F(c^* - 1|\mu, \beta) > F(c^*|\mu, \beta) < F(c^* + 1|\mu, \beta). \tag{26}$$

Some examples are performed to illustrate the existence of solution. We set $\mu = 20$, $\beta = 0.8$ and consider the following cost parameters as

$$C_h = \$150/\text{customer}/\text{unit time}, \quad C_d = \$30/\text{server}/\text{unit time}, \quad C_r = \$45/\text{unit time},$$
$$C_s = \$15/\text{unit time}, \quad C_l = \$120/\text{unit time}, \quad C_p = \$60/\text{server}.$$

Under other parameters are given, we observe from Table 1 that (i) the optimal number of servers c^* and its corresponding minimum cost increase as λ or α increases; and (ii) the optimum number of servers, c^* does not affect at all when γ changes from 0.2 to 0.8. This seems too insensitive to changes in γ.

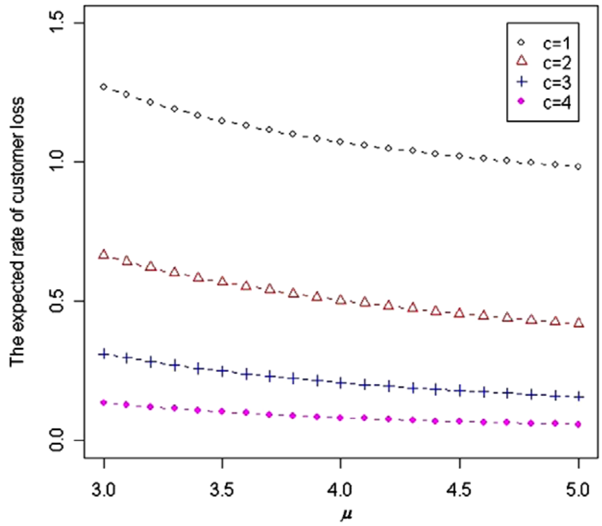

Fig. 10. The effect of μ on the expected loss rate.

Table 1
The cost function associated with a different number of servers and values of λ, α, γ.

$(\lambda, \mu, \alpha, \beta, \gamma)$	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
(5, 20, 0.05, 0.8, 0.2)	560.562	**519.969**	562.455	621.086	682.394	744.097
(10, 20, 0.05, 0.8, 0.2)	803.648	633.410	**619.825**	662.907	720.719	781.740
(15, 20, 0.05, 0.8, 0.2)	1133.07	821.187	**711.592**	716.238	762.168	820.114
(20, 20, 0.05, 0.8, 0.2)	N/A	1141.36	870.786	**797.168**	812.906	861.167
(5, 20, 0.05, 0.8, 0.5)	560.921	**520.018**	562.462	621.086	682.394	744.097
(10, 20, 0.05, 0.8, 0.5)	808.851	633.706	**619.871**	662.915	720.720	781.740
(15, 20, 0.05, 0.8, 0.5)	1142.56	821.620	**711.672**	716.257	762.172	820.115
(20, 20, 0.05, 0.8, 0.5)	N/A	1141.36	870.792	**797.170**	812.907	861.167
(5, 20, 0.05, 0.8, 0.8)	561.137	**520.064**	562.469	621.087	682.394	744.097
(10, 20, 0.05, 0.8, 0.8)	812.638	633.970	**619.915**	662.922	720.721	781.740
(15, 20, 0.05, 0.8, 0.8)	1149.01	822.028	**711.770**	716.276	762.176	820.116
(20, 20, 0.05, 0.8, 0.8)	N/A	1141.36	870.797	**797.172**	812.908	861.167
$(\lambda, \mu, \alpha, \beta, \gamma)$	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
(5, 20, 0.1, 0.8, 0.5)	589.183	**535.802**	571.222	628.410	690.473	753.557
(10, 20, 0.1, 0.8, 0.5)	847.825	668.532	**639.509**	674.618	730.219	791.597
(15, 20, 0.1, 0.8, 0.5)	1190.21	879.252	749.803	**738.112**	775.973	831.496
(20, 20, 0.1, 0.8, 0.5)	N/A	1220.18	934.448	837.518	**836.582**	876.824
(5, 20, 0.3, 0.8, 0.5)	676.654	**598.853**	611.031	658.310	718.775	784.148
(10, 20, 0.3, 0.8, 0.5)	968.223	797.030	**726.312**	731.023	771.576	828.080
(15, 20, 0.3, 0.8, 0.5)	1337.30	1079.67	903.843	**842.111**	845.522	882.825
(20, 20, 0.3, 0.8, 0.5)	N/A	1481.37	1171.85	1015.11	**957.632**	959.272
(5, 20, 0.5, 0.8, 0.5)	737.368	655.188	**653.066**	690.071	745.381	809.312
(10, 20, 0.5, 0.8, 0.5)	1051.49	904.517	813.597	**795.664**	819.693	866.317
(15, 20, 0.5, 0.8, 0.5)	1438.88	1238.18	1045.98	954.811	**930.227**	946.340
(20, 20, 0.5, 0.8, 0.5)	N/A	1677.72	1373.52	1191.21	1097.83	**1065.77**

* Denotes system is unstable (i.e., the stable condition does not hold).

5.2. Quasi-Newton method

After we find c^*, we will use Quasi-Newton method to search/adjust (μ, β) until the minimum value of $F(c^*, \mu, \beta)$ is achieved, say $F(c^*, \mu^*, \beta^*)$. The cost minimization problem can be illustrated mathematically as

$$F(c^*, \mu^*, \beta^*) = \underset{\mu,\beta \text{ s.t. Eq. (10)}}{\text{Minimize}}\ F(c^*, \mu, \beta). \tag{27}$$

The finding of the joint optimal value (μ^*, β^*) for a given c^* is difficult to implement. We note that the derivative of the cost function F with respect to (μ, β) indicates the direction which cost function increases. It means that, the optimal value (μ^*, β^*) can be found along this opposite direction of the gradient. (See [30]). An effective procedure that makes it possible to calculate the optimal value (μ^*, β^*) is presented as follows:

Table 2
The illustration of the implement process of the Quasi-Newton method.

Case (i): $(\lambda, \alpha, \gamma) = (15, 0.1, 0.5)$ with initial value $(\mu, \beta) = (20, 0.8)$

Iterations	0	1	2	3	4	5
$F(c^*, \mu, \beta)$	738.1120	730.1445	727.7439	727.6845	727.6845	727.6845
c^*	4	4	4	4	4	4
μ	20	15.55359	16.61114	16.81859	16.82467	**16.82467**
β	0.8	0.968384	0.985683	0.983808	0.983693	**0.983693**
$\frac{\partial F}{\partial \mu}$	5.294032	−4.03993	−0.566575	−0.01552	−0.000013	-6.8×10^{-12}
$\frac{\partial F}{\partial \beta}$	−12.3544	−6.33340	−0.525819	−0.00908	−0.000007	-1.8×10^{-11}
L_s	0.735673	0.931753	0.878269	0.868294	0.868004	0.868003
$E[D]$	0.444444	0.374397	0.368432	0.369069	0.369108	0.369108
L.R.	0.320231	0.518905	0.428566	0.415152	0.414787	0.414787
Hessian	$\begin{bmatrix} 1.3099 & 3.1502 \\ 3.1502 & 156.56 \end{bmatrix}$	$\begin{bmatrix} 3.7520 & 4.1618 \\ 4.1618 & 111.69 \end{bmatrix}$	$\begin{bmatrix} 2.7621 & 3.4215 \\ 3.4215 & 98.124 \end{bmatrix}$	$\begin{bmatrix} 2.6135 & 3.3275 \\ 3.3275 & 97.351 \end{bmatrix}$	$\begin{bmatrix} 2.6093 & 3.3252 \\ 3.3252 & 97.347 \end{bmatrix}$	$\begin{bmatrix} 2.6093 & 3.3251 \\ 3.3251 & 97.347 \end{bmatrix}$

Case (ii): $(\lambda, \alpha, \gamma) = (10, 0.3, 0.5)$ with initial value $(\mu, \beta) = (20, 0.8)$

Iterations	0	1	2	3	4	5	6
$F(c^*, \mu, \beta)$	726.3119	693.9217	673.7195	671.5941	671.5440	671.5439	671.5439
c^*	3	3	3	3	3	3	3
μ	20	12.2204	13.99813	14.58737	14.62913	14.62891	**14.62891**
β	0.8	1.13777	1.432544	1.600997	1.639305	1.640787	**1.640789**
$\frac{\partial F}{\partial \mu}$	6.46736	−10.0017	−2.16940	−0.178865	−0.00212	-1.7×10^{-6}	-1.8×10^{-12}
$\frac{\partial F}{\partial \beta}$	−123.276	−56.5653	−15.5898	−2.36610	−0.08194	−0.00011	-1.0×10^{-10}
L_s	0.480083	0.733697	0.666151	0.645710	0.644638	0.644667	0.644667
$E[D]$	0.818182	0.625970	0.519467	0.473436	0.464084	0.463729	0.463729
L.R.	0.946575	1.254851	0.864804	0.747327	0.711000	0.730524	0.730523
Hessian	$\begin{bmatrix} 0.9484 & 2.6956 \\ 2.6956 & 427.06 \end{bmatrix}$	$\begin{bmatrix} 5.0396 & 3.5377 \\ 3.5377 & 170.56 \end{bmatrix}$	$\begin{bmatrix} 3.0123 & 2.3415 \\ 2.3415 & 84.356 \end{bmatrix}$	$\begin{bmatrix} 2.5564 & 1.8817 \\ 1.8817 & 59.713 \end{bmatrix}$	$\begin{bmatrix} 2.5221 & 1.8093 \\ 1.8093 & 55.579 \end{bmatrix}$	$\begin{bmatrix} 2.5219 & 1.8069 \\ 1.8069 & 55.430 \end{bmatrix}$	$\begin{bmatrix} 2.5219 & 1.8069 \\ 1.8069 & 55.429 \end{bmatrix}$

Table 3
The optimal value (c^*, μ^*, β^*) and its minimum expected value for various values of λ, α, and γ, while c^* is obtained at initial value $(\mu, \beta) = (20, 0.8)$.

$(\lambda, \alpha, \gamma)$	(5, 0.1, 0.5)	(10, 0.1, 0.5)	(15, 0.1, 0.5)	(5, 0.5, 0.5)	(10, 0.5, 0.5)	(15, 0.5, 0.5)
c^*	2	3	4	3	4	5
(μ^*, β^*)	[9.313793, 0.818691]	[14.02913, 0.951822]	[16.82467, 0.983693]	[8.896326,1.510552]	[13.41282, 1.833732]	[16.42465, 1.968097]
$F(c^*, \mu^*, \beta^*)$	449.8870	608.4378	727.6845	540.4650	715.4502	849.7029
L_s	0.493536	0.679731	0.868003	0.530139	0.716304	0.889476
$E[D]$	0.217701	0.285219	0.369108	0.746064	0.856996	1.012926
L.R.	0.606480	0.538771	0.414787	0.476188	0.488205	0.424664
$(\lambda, \alpha, \gamma)$	(15, 0.1, 0.2)	(15, 0.3, 0.2)	(15, 0.5, 0.2)	(15, 0.1, 0.8)	(15, 0.3, 0.8)	(15, 0.5, 0.8)
c^*	4	4	5	4	4	5
(μ^*, β^*)	[16.82713, 0.983389]	[17.72648, 1.720770]	[16.42762, 1.967314]	[16.82225, 0.983991]	[17.72113, 1.722408]	[16.42172, 1.968856]
$F(c^*, \mu^*, \beta^*)$	727.6578	793.3554	849.6615	727.7107	793.4822	849.7433
L_s	0.867909	0.816563	0.883965	0.868097	0.816661	0.889586
$E[D]$	0.369212	0.593833	1.013248	0.369007	0.593352	1.012615
L.R.	0.414464	0.581035	0.424300	0.415105	0.582143	0.425021
$(\lambda, \alpha, \gamma)$	(10, 0.1, 0.2)	(10, 0.1, 0.5)	(10, 0.1, 0.8)	(10, 0.5, 0.2)	(10, 0.5, 0.5)	(10, 0.5, 0.8)
c^*	3	3	3	4	4	4
(μ^*, β^*)	[14.02822, 0.951169]	[14.02913, 0.951822]	[14.03003, 0.952454]	[13.41353, 1.831961]	[13.41282, 1.833732]	[13.41214, 1.835407]
$F(c^*, \mu^*, \beta^*)$	608.3219	608.4378	608.5506	715.3165	715.4502	715.5793
L_s	0.679970	0.679731	0.679507	0.716492	0.716304	0.716125
$E[D]$	0.285397	0.285219	0.285048	0.857647	0.856996	0.856382
L.R.	0.537821	0.538771	0.539685	0.487267	0.488205	0.489114

Algorithm: Quasi-Newton Method

INPUT Cost function $F(c^*, \mu, \beta)$, $R, \lambda, \mu, \alpha, \beta, \gamma$, initial value $\overrightarrow{\theta}^{(0)} = [\mu^{(0)}, \beta^{(0)}]^{\mathrm{T}}$, and the tolerance ε.
OUTPUT approximation solution $[\mu^*, \beta^*]^{\mathrm{T}}$.

Step 1 Set the initial trial solution for $\overrightarrow{\theta}^{(0)}$, and compute $F(c^*, \mu^{(0)}, \beta^{(0)})$.
Step 2 While $|\partial F/\partial\mu| > \varepsilon$ or $|\partial F/\partial\beta| > \varepsilon$ do Steps 3–4
    Step 3 Compute the cost gradient $\vec{\nabla}F(\overrightarrow{\theta}) = [\partial F/\partial\mu, \partial F/\partial\beta]^{\mathrm{T}}$ and the cost Hessian matrix
    $H(\overrightarrow{\theta}) = \begin{bmatrix} \partial^2 F/\partial\mu^2 & \partial^2 F/\partial\mu\partial\beta \\ \partial^2 F/\partial\beta\partial\mu & \partial^2 F/\partial\beta^2 \end{bmatrix}$ at point $\overrightarrow{\theta}^{(i)}$.
    Step 4 Find the new trial solution $\overrightarrow{\theta}^{(i+1)} = \overrightarrow{\theta}^{(i)} - [H(\overrightarrow{\theta}^{(i)})]^{-1}\vec{\nabla}F(\overrightarrow{\theta}^{(i)})$.
Step 5 OUTPUT

We present two examples to illustrate the optimization procedure shown in Table 2. From Table 2, we can see that the minimum expected cost per day of **727.6845** is achieved at $(\mu^*, \beta^*) = (\mathbf{16.82467}, \mathbf{0.983693})$ by using 5 iterations, which is $c^* = 4$ based on Case (i) with initial value $(\mu, \beta) = (20, 0.8)$. Based on Case (ii), c^* is 3 and the minimum expected cost per day of **671.5439** is achieved at $(\mu^*, \beta^*) = (\mathbf{14.62891}, \mathbf{1.640789})$ by using 6 iterations.

Finally, we perform a sensitivity investigation to the optimal values (c^*, μ, β). For various values of λ, α, and γ by considering the initial value (μ, β) of (20, 0.8), the minimum expected cost $F(c^*, \mu, \beta)$ and the system performance measures L_s, $E[D]$ and L.R. at the optimum values (c^*, μ, β) are shown in Table 3.

From Table 3, we find that (i) c^* increases as λ or α increases and is insensitive to the change of γ; (ii) $\mu^*(\beta^*)$ increases as $\lambda(\alpha)$ increases. (iii) μ^* and β^* slightly changes when γ changes from 0.2 to 0.8. Intuitively, this seems too insensitive to changes in γ.

6. Concluding remarks

An infinite capacity M/M/c queueing system with balking, reneging and server breakdowns is studied. This system is formulated as a QBD process and the necessary and sufficient condition for the stability of the system is discussed. It generalizes the model studied by Wang and Chang [13]. We have not only obtained numerically the steady-state probability and the system performance measures using matrix approach but also presented one efficient method to find the optimal number of servers, the optimal service rate and repair rate, simultaneously, so as to reach the minimum cost. We also have performed a sensitivity analysis between the joint optimal values (c^*, μ^*, β^*) and specific values of λ, α, and γ.

Acknowledgements

The authors would like to express their sincere thanks to the referees and the editor for their helpful suggestions that greatly improved the quality of this paper.

References

[1] C.J. Ancker Jr., A.V. Gafarian, Some queueing problems with balking and reneging: I, Oper. Res. 11 (1963) 88–100.
[2] C.J. Ancker Jr., A.V. Gafarian, Some queueing problems with balking and reneging: II, Oper. Res. 11 (1963) 928–937.
[3] J.-C. Ke, K.-H. Wang, Cost analysis of the M/M/R machine repair problem with balking, reneging, and server breakdowns, J. Oper. Res. Soc. 50 (1999) 275–282.
[4] K.-H. Wang, J.-B. Ke, J.-C. Ke, Profit analysis of the M/M/R machine repair problem with balking, reneging, and standby switching failures, Comput. Oper. Res. 34 (2) (2007) 835–847.
[5] F.A. Haight, Queueing with balking, Biometrika 44 (1957) 360–369.
[6] F.A. Haight, Queueing with reneging, Metrika 2 (1959) 186–197.
[7] M.O. Abou-El-Ata, A.I. Shawky, The single-server Markovian overflow queue with balking, reneging and an additional server for longer queues, Microelectron. Reliab. 32 (1992) 1389–1394.
[8] M.O. Abou-El-Ata, A.M.A. Hariri, The M/M/c/N queue with balking and reneging, Comput. Oper. Res. 19 (1992) 713–716.
[9] R.O. Al-Seedy, The truncated queue: M/M/2/m/m + Y with balking, spares, machine interference and an additional server for longer queues (Krishnamoorthi discipline), Microelectron. Reliab. 35 (1995) 1423–1427.
[10] R.O. Al-Seedy, Analytical solution of the state-dependent Erlangian queue: M/Ej/1/N with balking, Microelectron. Reliab. 36 (1996) 203–206.
[11] A.I. Shawky, The single-server machine interference model with balking reneging and an additional server for longer queues, Microelectron. Reliab. 37 (1997) 355–357.
[12] K.-H. Wang, J.-C. Ke, Probabilistic analysis of a repairable system with warm standbys plus balking and reneging, Appl. Math. Model. 27 (2003) 327–336.
[13] K.-H. Wang, Y.-C. Chang, Cost analysis of a finite M/M/R queueing system with balking, reneging, and server breakdowns, Math. Methods Oper. Res. 56 (2002) 169–180.
[14] R.O. Al-Seedy, A transient solution of the non-truncated queue M/M/2 with balking and an additional server for longer queues (Krishnamoorthi discipline), Appl. Math. Comput. 156 (3) (2004) 763–769.
[15] K.-H. Wang, S.-C. Chen, J.-C. Ke, Maximum likelihood estimates and confidence intervals of an M/M/R/N queue with balking and heterogeneous servers, RAIRO Oper. Res. 38 (2004) 227–242.
[16] J.-C. Ke, Operating characteristic analysis on the $M^{[x]}/G/1$ system with a variant vacation policy and balking, Appl. Math. Model. 31 (7) (2007) 1321–1337.
[17] D. Yue, W. Yue, Analysis of M/M/c/N queueing systems with balking, reneging and synchronous vacations, in: The 2nd Asia-Pacific Symposium on Queueing Theory and Network Applications, QTNA2007, pp. 53–62.

[18] D.P. Gaver, A waiting line with interrupted service, including priorities, J. R. Stat. Soc. Ser. B 24 (1962) 73–96.
[19] Y.H. Tang, A single-server M/G/1 queueing system subject to breakdowns—some reliability and queueing problem, Microelectron. Reliab. 37 (2) (1997) 315–321.
[20] W. Li, D. Shi, X. Chao, Reliability analysis of M/G/1 queueing systems with server breakdowns and vacations, J. Appl. Probab. 34 (1997) 546–555.
[21] J. Wang, An M/G/1 queue with second optional service and server breakdowns, Comput. Math. Appl. 47 (2004) 1713–1723.
[22] K.-H. Wang, M.-C. Chan, J.-C. Ke, Maximum entropy analysis of the $M^{[x]}/M/1$ queueing system with multiple vacations and server breakdowns, Comput. Ind. Eng. 52 (2) (2007) 192–202.
[23] G. Choudhury, L. Tadj, An M/G/1 queue with two phases of service subject to the server breakdown and delayed repair, Appl. Math. Model. 33 (2009) 2699–2709.
[24] G. Choudhury, K. Deka, An M/G/1 retrial queueing system with two phases of service subject to the server breakdown and repair, Perform. Eval. 65 (10) (2008) 714–724.
[25] L. Tadj, G. Choudhury, Optimal design and control of queues, TOP 13 (2005) 359–414.
[26] J.-C. Ke, Optimal NT policies for M/G/1 system with a startup and unreliable server, Comput. Ind. Eng. 50 (3) (2006) 248–262.
[27] K.-H. Wang, T.-Y. Wang, W.L. Pearn, Optimal control of the N policy M/G/1 queueing system with server breakdowns and general startup times, Appl. Math. Model. 31 (10) (2007) 2199–2212.
[28] M.F. Neuts, Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach, The John Hopkins University Press, Baltimore, 1981.
[29] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, in: ASA–SIAM Series on Statistics and Applied Probability, 1999.
[30] E.K.P. Chong, S.H. Zak, An Introduction to Optimization, 2nd ed., John Wiley and Sons, Inc., 2001.