Int J Adv Manuf Technol (2007) 33: 42–47
DOI 10.1007/s00170-006-0450-5

ORIGINAL ARTICLE

Kuo-Hsiung Wang · Jyh-Bin Ke · W. L. Pearn

Optimal management for a finite M/M/R queueing system with two arrival modes

Received: 4 September 2005 / Accepted: 30 January 2006 / Published online: 30 March 2006

Abstract We consider an M/M/R queueing system with finite capacity N, where customers have two arrival modes under steady-state conditions. It is assumed that each arrival mode is serviced by one or more servers, and that the two arrival modes have equal probabilities of receiving service. Arrival times of the customers and service times of the severs follow an exponential distribution. A cost model is developed to determine the optimal number of servers and the optimal system capacity. The minimum expected cost, the optimal number of servers, the optimal system capacity, and various system characteristics are obtained for some designated system parameters' values. Sensitivity for the minimal cost is also investigated.

Keywords Cost · Optimization · Queue · Sensitivity analysis · Two arrival modes

1 Introduction

This paper considers an M/M/R queueing system with finite capacity N, where N is the maximum number of customers in the system. We assume that the customers have two arrival modes, and are serviced by one or more servers in the service facilities. We also assume that any one mode of arrival can be serviced by one or more servers, and that each arrival mode has an equal probability of receiving service.

Each customer has two independent arrival modes (mode 1 and mode 2). Both arrival modes 1 and 2 of customers follow a Poisson process with parameters λ_1 and λ_2, respectively. Suppose that both modes are equally likely to be serviced next, when several customers are waiting for service. The service time of each server in the service mode i has an exponential distribution with mean $1/\mu_i$, where i=1, 2. The arriving customers join in a single waiting line based on the order of their arrivals; that is, in a first-come, first-served discipline. Each server services only one customer at a time. Customers who, upon entry into the service facility, find that the server is busy have to wait in the queue until the server is available.

Analytic steady-state solutions of a finite-capacity M/M/R queueing system with two arrival modes have not been found. For cases with one single arrival mode, analytic steady- state solutions of an M/M/R queueing system have been provided by several authors, including Gross and Harris [1] and Kleinrock [2]. The past work for the two (or multiple) arrival modes situation may be divided into two parts, according to whether the queueing model is of infinite source or finite source. In the first category, we review a previous paper which deals with infinite source queues. Tijms [3] considered a finite-capacity queueing system with two arrival modes (which appeared as an exercise, p. 152, Exercise 2.23), but provided no cost analysis or solution methodologies. The second category of authors deals with papers treating the finite source model. Finite source models as applied to machine repair problems have been examined by several researchers. Without deriving analytic steady-state solutions, Benson and Cox [4] first considered the no-spare M/M/1 machine repair problem with two failure modes. Again, without providing analytic steady-state solutions, Elsayed [5] studied two repair policies for the no-spare M/M/1 machine repair problem with two failure modes. Analytic steady-state solutions of the cold-standby M/M/R machine repair problem with two failure modes were first derived by Wang [6]. Later on, Wang and Wu [7] investigated the M/M/R machine repair problem with spares where machines have two failure modes. Spares are considered to be either cold-standby, warm-standby, or hot-standby. Wang and Lee [8] extended Wang's model [6] to the M/M/R machine repair problem with multiple failure modes.

K.-H. Wang (✉) · J.-B. Ke
Department of Applied Mathematics,
National Chung-Hsing University,
Taichung, Taiwan, 402,
Republic of China
e-mail: khwang@amath.nchu.edu.tw

W. L. Pearn
Department of Industrial Engineering and Management,
National Chiao Tung University,
Hsinchu, Taiwan, Republic of China

The birth-and-death process is used to derive analytic steady-state solutions to a finite capacity M/M/R queueing system with two arrival modes. This paper differs from past works in that: (a) the infinite-source queueing problem has distinct characteristics which are different from the machine repair problem; (b) it studies a finite-capacity M/M/R queueing system with two arrival modes; and (c) it performs a sensitivity analysis for the total expected cost with respect to specific values of the system parameters.

As an application of that problem, we consider a parking lot problem where there are two different classes of parkers, namely, long-period parkers and short-period parkers, arriving at a parking lot according to independent Poisson processes with rates λ_1 and λ_2 per hour, respectively. The parking lot has space for N cars. There are R employees in the parking lot management office. The service times of those employees for each class of parkers follow an exponential distribution with mean $1/\mu_i$, where i=1, 2, respectively. The manager of the parking lot management office would like to know the system performance for the parking lot, such as the expected number of cars in the parking lot, the expected number of the busy employees, and the expected number of idle employees, for minimum cost.

We first develop analytic steady-state solutions for the finite M/M/R queueing system with two arrival modes by using the modified birth-and-death results. Next, a cost model is developed to determine the optimal values of the number of servers and the system capacity, simultaneously, in order to minimize the total expected cost per unit time. Finally, various system performance measures are evaluated under optimal operating conditions. We perform a sensitivity analysis for the minimal cost with respect to changes in specific values of some system parameters.

2 Steady-state results

The system can be analyzed as a continuous time parameter Markov chain with states $\{(i, j)|i+j=0,\ 1,\ 2, \ldots,\ N\}$, where i denotes the number of customers of mode 1 and j is the number of customers of mode 2. For a steady-state condition, let $P(i, j)$≡probability that there are i and j customers of modes 1 and 2 in the system, respectively. The mean arrival rate, $\lambda_{i,j}^{k}$, for arrival mode k (k=1, 2) are as follows:

$$\lambda_{i,j}^{1} = \begin{cases} \lambda_1 & \text{if } 0 \le i+j < N \\ 0 & \text{otherwise} \end{cases}$$

$$\lambda_{i,j}^{2} = \begin{cases} \lambda_2 & \text{if } 0 \le i+j < N \\ 0 & \text{otherwise} \end{cases}$$

Let n represent the number of customers of mode 1 (or mode 2) in the system. The mean service rate, μ_n^{k}, for service mode k (k=1, 2) are as follows:

$$\mu_n^{1} = \begin{cases} n\mu_1 & \text{if } 0 \le n \le R \\ R\mu_1 & \text{if } R \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_n^{2} = \begin{cases} n\mu_2 & \text{if } 0 \le n \le R \\ R\mu_2 & \text{if } R \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

A special case of a finite-capacity M/M/R queueing system with arrival modes is shown in Fig. 1. The steady-state

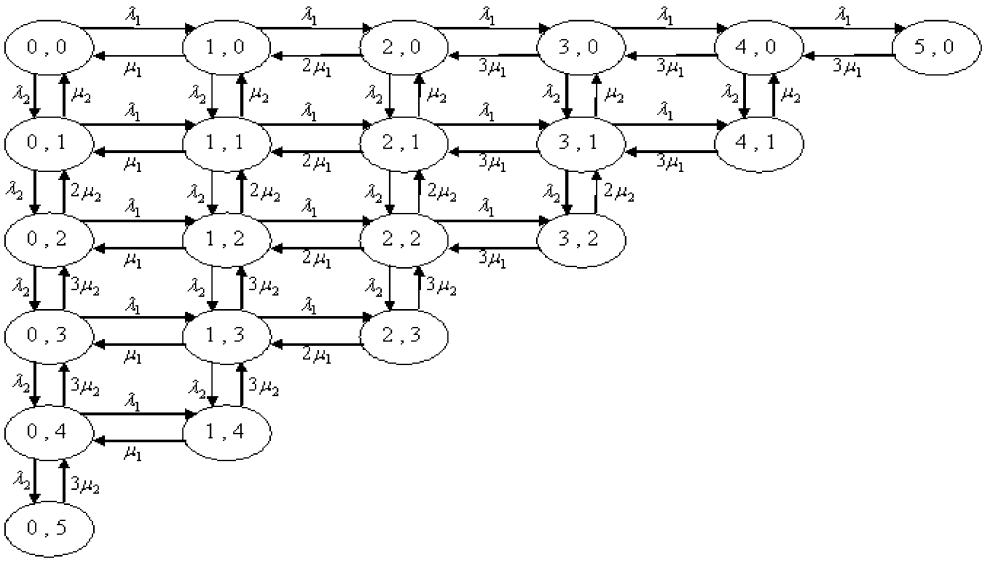

Fig. 1 State-transition-rate diagram for an M/M/R/N queueing system with two arrival modes (R=3, N=5)

equations for $P(i, j)$ for an M/M/R queueing system with finite capacity N and two arrival modes are given by:

$$(\lambda_1 + \lambda_2)P(0,\ 0) = \mu_1 P(1,\ 0) + \mu_2 P(0,\ 1) \tag{1}$$

$$\begin{aligned}&[\lambda_1 + \lambda_2 + \min(i,\ R)\mu_1]P(i,\ 0)\\ &= \lambda_1 P(i-1,\ 0) + \mu_2 P(i,\ 1)\\ &\quad + \min(i+1,\ R)\mu_1 P(i+1,\ 0)\\ &\quad 1 \le i \le N-1\end{aligned} \tag{2}$$

$$R\mu_1 P(N,\ 0) = \lambda_1 P(N-1,\ 0) \tag{3}$$

$$\begin{aligned}&[\lambda_1 + \lambda_2 + \min(j,\ R)\mu_2]P(0,\ j)\\ &= \lambda_2 P(0,\ j-1) + \mu_1 P(1,\ j)\\ &\quad + \min(j+1,\ R)\mu_2 P(0,\ j+1)\\ &\quad 1 \le j \le N-1\end{aligned} \tag{4}$$

$$R\mu_2 P(0,\ N) = \lambda_2 P(0,\ N-1) \tag{5}$$

$$\begin{aligned}&[\lambda_1 + \lambda_2 + \min(i,\ R)\mu_1 + \min(j,\ R)\mu_2]P(i,\ j)\\ &= \lambda_1 P(i-1,\ j) + \lambda_2 P(i,\ j-1)\\ &\quad + \min(i+1,\ R)\mu_1 P(i+1,\ j)\\ &\quad + \min(j+1,\ R)\mu_2 P(i,\ j+1)\\ &\quad 1 \le i,\ j \le N-1,\ 2 \le i+j \le N-1\end{aligned} \tag{6}$$

$$\begin{aligned}&[\min(i,\ R)\mu_1 + \min(N-i,\ R)\mu_2]P(i,\ N-1)\\ &= \lambda_1 P(i-1,\ N-1) + \lambda_2 P(i,\ N-i-1)\\ &1 \le i \le N-1\end{aligned} \tag{7}$$

Solving Eqs. 1, 2, 3, 4, 5, 6, 7 recursively or by using the following known formula, which can be found in [6, 7]:

$$P(i,\ j) = \left[\prod_{n=1}^{i} \frac{\lambda^1_{n-1,\ 0}}{\mu^1_n}\right]\left[\prod_{n=1}^{j} \frac{\lambda^2_{i,\ n-1}}{\mu^2_n}\right] P(0,\ 0) \tag{8}$$

where $a>b$ in the $\prod_{l=a}^{b}(\cdot)$ notation indicates that the term is 1, we obtain respectively:

$$P(i,\ 0) = \frac{\rho_1^i}{\prod_{n=1}^{i} \min\,(n,\ R)} P(0,\ 0) \quad 1 \le i \le N \tag{9}$$

$$P(0,\ j) = \frac{\rho_2^j}{\prod_{n=1}^{j} \min\,(n,\ R)} P(0,\ 0) \quad 1 \le j \le N \tag{10}$$

$$\begin{aligned}P(i,\ j) = &\frac{\rho_1^i \rho_2^j}{\prod_{n=1}^{i} \min(n,\ R) \prod_{n=1}^{j} \min(n,\ R)} P(0,\ 0)\\ &1 \le i+j \le N \quad i,\ j = 0,\ 1,\ 2, \ldots,\ N\end{aligned} \tag{11}$$

where $\rho_1 = \frac{\lambda_1}{\mu_1}$, $\rho_2 = \frac{\lambda_2}{\mu_2}$, and:

$$\prod_{l=1}^{n} \min(l,\ R) = \begin{cases} n! & \text{if } 1 \le n \le R \\ R!R^{n-R} & \text{if } R+1 \le n \le N \end{cases}$$

It should be noted that Eq. 9 (or Eq. 10) is identical to the results for a finite-capacity M/M/R queueing system with a single arrival mode (see Gross and Harris [1], p. 93). The steady-state solutions $P(i, j)$ always exist because the number of states is finite. We use an efficient Matlab computer program to evaluate $P(i, j)$ by using Eqs. 9, 10, 11 and the following normalizing equation:

$$\sum_{i=0}^{N} \sum_{j=0}^{N-i} P(i,\ j) = 1 \tag{12}$$

3 System performance measures

Our analysis is based on the following system performance measures of a finite M/M/R queueing system with two arrival modes. Let us note the following:

L_1	Expected number of customers in the system for arrival mode 1
L_2	Expected number of customers in the system for arrival mode 2
L_s	Expected number of customers in the system
L_q	Expected number of customers in the queue
$E[I]$	Expected number of idle servers
$E[B]$	Expected number of busy servers

The expressions for L_1, L_2, L_s, L_q, $E[I]$, and $E[B]$ are given by:

$$L_1 = \sum_{i=1}^{N} i \left[\sum_{j=0}^{N-i} P(i,\ j) \right] \tag{13}$$

$$L_2 = \sum_{j=1}^{N} j \left[\sum_{j=0}^{N-j} P(i,\ j) \right] \tag{14}$$

$$L_s = \sum_{i+j=0}^{N} (i+j) P(i,\ j) = L_1 + L_2 \tag{15}$$

$$\begin{aligned} L_q = & \sum_{i=R+1}^{N} (i-R) \sum_{j=0}^{N-R-1} P(i,\ j) \\ & + \sum_{j=R+1}^{N} (j-R) \sum_{i=0}^{N-R-1} P(i,\ j) \qquad N > R \end{aligned} \tag{16}$$

$$E[I] = \sum_{i+j=0}^{R-1} [R - \max(i,\ j)] P(i,\ j) \tag{17}$$

$$E[B] = R - E[I] \tag{18}$$

4 Cost sensitivity analysis

We develop a total expected cost function per unit time for an M/M/R queueing system with finite capacity N and two arrival modes, in which R and N are two decision variables. Our objective is to determine the optimum number of servers R, say R^*, and the optimal system capacity N, say N^*, simultaneously, so as to minimize this function. Let C_1 be the holding cost per unit time per customer present in the system, C_2 be the cost per unit time when one server is idle, C_3 be the cost per unit time when one server is busy, C_4 be the fixed cost for every customer's space, and C_4 be the fixed cost for every lost customer. Thus, the total expected cost function per unit time is given by:

$$\begin{aligned} & F(R,\ N) \\ & \quad = C_1 L_s + C_2 E[I] + C_3 E[B] + C_4 N \\ & \qquad + (\lambda_1 + \lambda_2) C_5 P_N \end{aligned} \tag{19}$$

where $P_N = \sum_{i+j+N} P(i,\ j)$.

The cost parameters in Eq. 19 are assumed to be linear in the expected number of the indicated quantity. Substitution of Eqs. 11 and 15, 16, 17, 18 into Eq. 19, the cost function $F(R, N)$ is too detailed to be shown here. Hence, it would have been an arduous task or, at least, extremely difficult to develop the optimal solution (R^*, N^*) symbolically, due to the highly non-linear and complex nature of the optimization problem. To the best of the authors' knowledge, no new and efficient methods to solve this optimization problem currently exist. This is due to the fact that there are two decision variables, R and N, involved in our model. Here, we should point out explicitly that the solution really gives the minimum value. Therefore, we will perform the numerical experiments to show that the cost function is really convex and that the solution gives a minimum. An efficient and direct procedure is used to obtain (R^*, N^*). Following Hilliard [9], we carry out the following steps for achieving the optimal value (R^*, N^*):

Step 1
Find the optimal system capacity N^*, for R servers, i.e., $\min_N F(R,\ N) = F(R,\ N^*)$

Step 2
Find the set of all minimum cost solutions for R=1, 2,…, N, i.e., $\Theta = \{F(R,\ N^*) :\ R = 1,\ 2, \ldots,\ N\}$

Step 3
Find the optimal number of servers, R^*, i.e., $\min_R \Theta = F(R^*,\ N^*)$.

The following numerical results are obtained for C_1=\$8/h, C_2=\$10/h, C_3=\$30/h, C_4=\$5/h, and C_5=\$25/h. We fix λ_1=20 arrivals/h, μ_0=10 services/h, λ_2=10 arrivals/h, μ_2=10 services/h, vary the number of servers R from 1 to 6, and vary the system capacity N from 3 to 12. The expected cost $F(R, N)$ is shown in Table 1 for various values of R and N. We note that a minimum expected cost per hour of \$160.54 is obtained with R^*=4 and N^*=9.

To find (R^*, N^*), we should show the existence of convexity or unimodality of $F(R, N)$. However, this task is difficult to implement. The function $F(R, N)$ is unimodal; that is, it has a single relative minimum. The numerical results shown in Table 1 can convince us that the cost function is convex. The minimum expected cost $F(R, N)$, the values of various system characteristics L_s, $E[I]$, $E[B]$, and the sensitivities of $F(R, N)$ with respect to λ_1, μ_1, λ_2, and μ_2, at the optimal value (R^*, N^*) are shown in Table 2 for (λ_2, μ_2)=(20, 10) and different values of (λ_1, μ_1). From Table 2, as would be expected, we observe that: (1) $F(R^*, N^*)$ increases as λ_1 increases or μ_1 decreases, which can be easily predicted by the sign of its sensitivity; and (2) the optimum values of (R, N), (R^*, N^*) increases as λ_1 increases. From the last three columns of Table 2, R^* does not change and N^* rarely changes when μ_1 changes from 15 to 30. Intuitively, this seems to be too insensitive to changes in μ_1. This phenomenon also can be seen from the order of sensitivities. (i.e., $\partial F / \partial \mu_1$ is the smallest in the last three columns of Table 2).

The sensitivity analysis results are shown in Table 3 by increasing λ_1 from 0.01 to 10,000 and fixing μ_1=20 and (λ_2, μ_2)=(20, 10). We observe from Table 3 that: (1) as λ_1 increases, the impact of λ_1 on $F(R, N)$ increases and

Table 1 The expected cost $F(R, N)$ for $(\lambda_1, \mu_1)=(20, 10)$, $(\lambda_2, \mu_2)=(10, 10)$

N	R					
	1	2	3	4	5	6
3	496.00	367.33	351.84	–	–	–
4	483.65	311.26	268.80	270.68	–	–
5	481.98	277.71	218.65	212.29	219.89	–
6	486.45	257.48	190.24	180.33	185.78	195.14
7	494.51	245.52	174.91	165.64	170.66	179.52
8	504.67	238.91	167.35	160.59	166.37	175.23
9	516.05	235.91	164.50	*160.54*	167.39	176.51
10	528.13	235.44	164.55	163.03	170.74	180.14
11	540.62	236.80	166.39	166.79	175.08	184.66
12	553.34	239.50	169.37	171.18	179.81	189.50

Table 3. Sensitivity analysis of $F(R, N)$ evaluated at $(R, N)=(4, 8)$, $\mu_1=10$, $(\lambda_2, \mu_2)=(20, 10)$ and varying λ_1

λ_1	$F(4, 8)$	$\partial F/\partial\lambda_1$	$\partial F/\partial\mu_1$	$\partial F/\partial\lambda_2$	$\partial F/\partial\mu_2$
0.01	139.6	1.42	−0.001	3.87	−7.46
1	141.0	1.53	−0.14	3.92	−7.57
5	148.1	2.04	−0.91	4.29	−8.12
10	160.6	3.02	−2.64	4.95	−9.41
15	179.4	4.64	−6.00	5.91	−10.52
30	311.7	13.8	−33.55	9.47	−13.70
100	2,029	26.6	−115.3	15.75	−1.32
1,000	24,703	25	−102	24	0
10,000	249,721	25	−100	25	0

eventually reaches a stable level; (2) a similar trend also occurs for the impacts of μ_1 and λ_2 on $F(R, N)$, but not μ_2; (3) the system with smaller λ_1 has a light traffic at mode 1, therefore, μ_1 does not affect $F(R, N)$; (4) the system with larger λ_1 has a bottleneck at mode 1, therefore, μ_2 does not affect $F(R, N)$ and the values of λ_1 and λ_2 have the same impact on $F(R, N)$; and (5) the most dominant parameter is μ_2 for smaller λ_1 and shifts to μ_1 for larger λ_1. (That is, we can significantly reduce the system cost by increasing the values of these dominant parameters.)

Next, the sensitivity analysis results are shown in Table 4 by increasing μ_1 from 0.01 to 10,000 and fixing $\lambda_1=15$ and $(\lambda_2, \mu_2)=(20, 10)$. We observe from Table 4 that: (1) as μ_1 increases, the impact of μ_1 on $F(R, N)$ decreases and reaches to zero eventually; (2) the similar trend also happens to the impacts of λ_1 and λ_2 on $F(R, N)$, but not μ_2; (3) the system with smaller μ_1 has a bottleneck at mode 1, therefore, μ_2 does not affect $F(R, N)$; (4) the system with larger μ_1 has a bottleneck at mode 2, therefore, μ_1 itself does not affect $F(R, N)$; and (5) the most dominant parameter is μ_1 for smaller μ_1 and shifts to μ_2 for larger μ_1. (That is, we can significantly reduce the system cost by increasing the value of these dominant parameters.)

5 Conclusions

In this paper, we modeled a finite M/M/R queueing system with two arrival modes, and obtained the steady-state analytic solutions. The considered model generalizes the existing M/M/R queueing system with infinite capacity and two arrival modes, the M/M/R queueing system with infinite capacity and a single arrival mode, and the finite M/M/R queueing system with a single arrival mode. We have provided an efficient method to determine the optimal number of servers and the optimal system capacity simultaneously, in order to minimize the expected cost function, and evaluated various system performance measures under the optimal operating conditions. We also performed a sensitivity analysis for the minimal expected cost with respect to specific values of λ_1, μ_1, λ_2, and μ_2. The results are useful for modeling banking service systems, computer jobs processing, performance evalua-

Table 2 System performance measures and sensitivities of a finite M/M/R queueing system with two arrival modes under optimal operating conditions for $(\lambda_2, \mu_2)=(20, 10)$ and various values of (λ_1, μ_1)

(λ_1, μ_1)	(30, 10)	(20, 10)	(15, 10)	(15, 15)	(15, 20)	(15, 30)
(R^*, N^*)	(5, 12)	(4, 11)	(4, 10)	(4, 9)	(4, 8)	(4, 8)
$F(R^*, N^*)$	233.43	191.32	174.58	161.50	155.94	150.39
L_s	5.241	4.260	3.657	3.121	2.843	2.608
$E[I]$	1.608	1.291	1.543	1.757	1.852	1.920
$E[B]$	3.392	2.709	2.457	2.243	2.148	2.080
$\partial F/\partial\lambda_1$	4.96	3.59	2.76	1.68	1.50	0.91
$\partial F/\partial\mu_1$	−14.06	−6.77	−3.87	−1.49	−0.90	−0.34
$\partial F/\partial\lambda_2$	2.90	3.59	3.80	4.29	5.06	4.84
$\partial F/\partial\mu_2$	−5.25	−6.77	−7.25	−8.19	−9.43	−9.23

Table 4 Sensitivity analysis of $F(R, N)$ evaluated at $(R, N)=(4, 8)$, $\lambda_1=15$, $(\lambda_2, \mu_2)=(20, 10)$ and varying μ_1

μ_1	$F(4, 8)$	$\partial F/\partial\lambda_1$	$\partial F/\partial\mu_1$	$\partial F/\partial\lambda_2$	$\partial F/\partial\mu_2$
0.01	1,097	25.09	−235.5	24.93	0
0.1	1,075	25.90	−235.6	24.33	0
1	863.0	34.05	−235.6	18.37	−0.05
5	272.5	19.17	−49.65	7.70	−10.15
10	179.4	4.64	−6.00	5.91	−10.52
20	155.9	1.50	−0.90	5.06	−9.54
50	146.5	0.55	−0.11	4.68	−8.99
100	144.0	0.32	−0.02	4.57	−8.82
1,000	141.8	0.15	0	4.48	−8.68
10,000	141.6	0.14	0	4.47	−8.66

tions, parking lot services, automatic machine car wash services, and many related other applications.

References

1. Gross D, Harris CM (1998) Fundamentals of queueing theory, 3rd edn. Wiley, New York
2. Kleinrock L (1975) Queueing systems, vol. I: theory. Wiley, New York
3. Tijms HC (1986) Stochastic modeling and analysis: a computational approach. Wiley, New York
4. Benson F, Cox DR (1951) The productivity of machines requiring attention at random intervals. J Roy Stat Soc B 13:65–82
5. Elsayed EA (1981) An optimum repair policy for the machine interference problem. J Oper Res Soc 32:793–801
6. Wang K-H (1994) Profit analysis of the machine repair problem with cold standbys and two modes of failure. Microelectron Reliab 34:1635–1642
7. Wang K-H, Wu J-D (1995) Cost analysis of the M/M/R machine repair problem with spares and two modes of failure. J Oper Res Soc 46(6):783–790
8. Wang K-H, Lee H-C (1998) Cost analysis of the cold-standby M/M/R machine repair problem with multiple modes of failure. Microelectron Reliab 38(3):435–441
9. Hilliard JE (1976) An approach to cost analysis of maintenance float systems. IIE Trans 8:128–133