Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Applied Mathematical Modelling 31 (2007) 2199–2212
APPLIED MATHEMATICAL MODELLING
www.elsevier.com/locate/apm

Optimal control of the N policy M/G/1 queueing system with server breakdowns and general startup times

Kuo-Hsiung Wang [a,*], Tsung-Yin Wang [b], Wen Lea Pearn [b]

[a] *Department of Applied Mathematics, National Chung-Hsing University, Taichung 402, Taiwan*
[b] *Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu 30050, Taiwan*

Received 1 August 2005; received in revised form 1 May 2006; accepted 30 August 2006
Available online 27 October 2006

Abstract

This paper deals with an N policy M/G/1 queueing system with a single removable and unreliable server whose arrivals form a Poisson process. Service times, repair times, and startup times are assumed to be generally distributed. When the queue length reaches $N(N \geqslant 1)$, the server is immediately turned on but is temporarily unavailable to serve the waiting customers. The server needs a startup time before providing service until there are no customers in the system. We analyze various system performance measures and investigate some designated known expected cost function per unit time to determine the optimal threshold N at a minimum cost. Sensitivity analysis is also studied.

Keywords: Control; General repair time; General startup time; M/G/1 queue; Optimal policy

1. Introduction

We consider an N policy M/G/1 queueing system in which the server is typically subject to unpredictable breakdowns. It is assumed that arrivals of customers follow a Poisson process and the breakdown times of the server follow the negative exponential distribution. We assume that the service times, the repair times, and the startup times obey a general distribution. The term ‘removable server’ is just an abbreviation for the system of turning on and turning off the server, depending on the number of customers in the system. When the number of customers in the system reaches the threshold $N(N \geqslant 1)$, the server is immediately turned on but is temporarily unavailable to serve waiting customers. He requires for the preparatory work (i.e., begin startup) before starting service. Once the startup is terminated, the server immediately starts serving the waiting customers.

It is assumed that arrivals of customers follow a Poisson process with rate λ. The service times for a customer are independent and identically distributed (i.i.d.) random variables obeying an arbitrary distribution function $S(t)(t \geqslant 0)$ with a mean service time μ_S and a finite variance σ_S^2. The server is subject to breakdowns

* Corresponding author.
E-mail address: khwang@amath.nchu.edu.tw (K.-H. Wang).

0307-904X/$ - see front matter
doi:10.1016/j.apm.2006.08.016

at any time with Poisson breakdown rate α when he is working. When the server fails, he is immediately repaired at a repair facility, where the repair times are independent and identically distributed random variables having a general distribution function $R(t)(t \geqslant 0)$ with a mean repair time μ_R and a finite variance σ_R^2. Arriving customers form a single waiting line at a server based on the order of their arrivals. The server can serve only one customer at a time and the service is independent of the arrival process. A customer who arrives and finds the server busy or broken down must wait in the queue until a server is available. Although no service occurs during the repair period of the server, customers continue to arrive following a Poisson process. Furthermore, when the queue length reaches a specific level, denoted by N, the server is immediately turned on (i.e. begin startup) but is temporarily unavailable to serve the waiting customers. He needs a startup time with random length before starting service. Again, the startup times are independent and identically distributed random variables obeying a general distribution function $U(t)(t \geqslant 0)$ with a mean startup time μ_U and a finite variance σ_U^2. Once the startup is terminated, the server begins serving the waiting customers until the system becomes empty. Service is allowed to be interrupted if the server breaks down, and the server is immediately repaired. Once the server is repaired, he immediately returns to serve customers until there are no customers in the system.

For a reliable server, Yadin and Naor [1] first introduced the concept of controllable queueing system under N policy. The N policy M/G/1 queueing system was first studied by Heyman [2] and was investigated by such authors as Bell [3], Tijms [4], Wang and Ke [5], and others. Exact steady-state solutions of the N policy $M/E_k/1$ and $M/H_k/1$ queueing systems were first developed by Wang and Huang [6] and Wang and Yen [7], respectively. Exact steady-state solutions of the N policy M/M/1 queueing system with exponential startup times were first derived by Baker [8]. Borthakur et al. [9] extended Baker's model to general startup times. The N policy M/G/1 queueing system with startup times was investigated by several authors such as Medhi and Templeton [10], Takagi [11], Lee and Park [12], Krishna et al. [13], Hur and Paik [14], etc. For an unreliable server, exact steady-state solutions of the N policy M/M/1, $M/E_k/1$, $M/H_2/1$, and $M/H_k/1$ queueing systems were developed by Wang [15], Wang [16], Wang et al. [17], and Wang et al. [18], respectively. Wang and Ke [19] studied three control policies in an M/G/1 queueing system and demonstrated that in three control policies, the probability that the server is busy in the steady-state is equal to the traffic intensity. Recently, Ke [20] examined the N policy M/G/1 queueing system with server vacations, startup and breakdowns. Furthermore, Ke and Pearn [21] investigated the N policy M/M/1 queueing system with heterogeneous arrivals, in which the server is characterized by breakdowns and vacations. Analytical sensitivity analysis of the N policy M/G/1 queueing system is investigated by Pearn et al. [22]. Exact steady-state solutions of the N policy M/M/1 queueing system with exponential startup times were first developed by Wang [23].

In this paper, we first develop various system performance measures, such as the expected number of customers, the expected length of the turned-off, complete startup, busy, and breakdown periods in the N policy M/G/1 queueing system with server breakdowns and general startup times. Next, we construct the total expected cost function per unit time to determine the optimal threshold N numerically in order to minimize the cost function. In addition, sensitivity analysis and some numerical examples are also investigated.

2. Justification of practical applications

A number of practical problems arise which may be formulated as one in which the server meets unpredictable breakdowns and requires a startup time before providing service. Such models have potentially useful in practical production(manufacturing) systems. For example, in reflow work for printed circuit board (PCB) surface mount. Assume that PCB arrives according to a random process. For cost concern, it is desirable that the reflow machine begins operating whenever the number of PCB reaches a critical value N. It takes random time for warming up before the reflow machine starts working. Moreover, the reflow process may be interrupted when machine encounters unpredicted breakdowns. When reflow interruptions occur (breakdowns), it is emergently recovered with a random time. Another possible application is wire bonding in integrated circuit (IC) assembly. To save cost, it is desirable that the wire bonder begins operating whenever the number of unbounded IC reaches a critical value N. It requires a random time for setup before the wire bonder starts working. The bonding process may be interrupted when the bonder meets breakdowns. When bonding interruptions occur (breakdowns), it is emergently recovered.

3. System performance measures

The primary objective of this section is to develop the various system performance measures, such as (i) expected number of customers in the system; (ii) expected length of the turned-off period, the complete startup period, the busy period, and the breakdown period; (iii) expected length of the busy cycle; and (iv) the probability that the server is turned-off, startup, busy and broken down.

3.1. Expected number of customers in the system

Let H be a random variable representing the completion time of a customer, which includes both the service time of a customer and the repair time of a server. Applying the well-known formula for the p.g.f. of the number of customers in the ordinary M/G/1 queueing system with reliable server, the p.g.f. of the number of customers in ordinary M/G/1 queueing system with unreliable server is given by

$$G(z) = \frac{(1-\rho_{\mathrm{H}})(1-z)\bar{f}_{\mathrm{H}}(\lambda-\lambda z)}{\bar{f}_{\mathrm{H}}(\lambda-\lambda z)-z}, \tag{1}$$

where $\rho_{\mathrm{H}} = \lambda E[H]$. In addition, $E[H] = \mu_{\mathrm{S}}(1+\alpha\mu_{\mathrm{R}})$ and $E[H^2] = (1+\alpha\mu_{\mathrm{R}})^2(\mu_{\mathrm{S}}^2+\sigma_{\mathrm{S}}^2)+\alpha\mu_{\mathrm{S}}(\mu_{\mathrm{R}}^2+\sigma_{\mathrm{R}}^2)$. It is to be noted that the traffic intensity ρ_{H} is assumed to be less than unity. We note that expression (1) is obtained only by replacing service times by completion times in the formula of the ordinary M/G/1 queueing system with reliable server.

For the N policy M/G/1 queueing system with server breakdowns requiring startup time, we consider that the server is on 'extended vacation' during the turned-off period plus the startup period. Following the result of Medhi and Templeton [10], we obtain

$$G_N(z) = \frac{[1-W(z)]G(z)}{W'(1)(1-z)}, \tag{2}$$

where

$G_N(z) =$ the p.g.f. of number of customers in the N policy M/G/1 queueing system with server breakdowns and general startup times.
$W(z) =$ the p.g.f. of the number of customers that arrive during the turned-off period plus the startup period;
$=$ [the p.g.f. of the number of customers that arrive during the turned-off period] × [the p.g.f. of the number of customers that arrive during the startup period];
$= z^N\bar{f}_{\mathrm{U}}(\lambda-\lambda z)$, where $\bar{f}_{\mathrm{U}}(\cdot)$ is the LST of startup time.

We have $W'(z) = Nz^{N-1}\bar{f}_{\mathrm{U}}(\lambda-\lambda z)+z^N\bar{f}_{\mathrm{U}}^{(1)}(\lambda-\lambda z)(-\lambda)$. It follows that $W'(1) = N+\lambda\mu_{\mathrm{U}}$, where $\mu_{\mathrm{U}} = -\bar{f}_{\mathrm{U}}^{(1)}(0)$ is the mean startup time. Let $\rho_{\mathrm{U}} = \lambda\mu_{\mathrm{U}}$. From (1) and (2), we obtain

$$G_N(z) = \frac{[1-z^N\bar{f}_{\mathrm{U}}(\lambda-\lambda z)](1-\rho_{\mathrm{H}})\bar{f}_{\mathrm{H}}(\lambda-\lambda z)}{(N+\rho_{\mathrm{U}})[\bar{f}_{\mathrm{H}}(\lambda-\lambda z)-z]}.$$

Let L_N denote the expected number of customers in the N policy M/G/1 queueing system with server breakdowns and general startup times. Thus we have

$$L_N = G_N'(z)|_{z=1} = \frac{1}{N+\rho_{\mathrm{U}}}\left[\frac{N(N-1)}{2}+N\rho_{\mathrm{U}}+\frac{\lambda^2E[U^2]}{2}\right]+\lambda E[H]+\frac{\lambda^2E[H^2]}{2[1-\lambda E[H]]}. \tag{3}$$

3.2. Expected length of the turned-off, complete startup, busy, and breakdown periods

The turned-off period terminates when the Nth customer arrives in system. Since the complete startup period starts when the turned-off period terminates, the complete startup period is represented by the sum of the startup period and the complete period. The server begins startup when there are at least N waiting customers in the system. This is called the startup period. The startup period terminates when the server starts to serve the waiting customers. Since the complete period begins when the startup period is over and terminates when the system becomes empty, the complete period is represented by the sum of the busy period and the breakdown period. The busy period is initiated when the server completes his startup and begins serving the waiting customers. During the busy period, the server may break down and starts his repair immediately. This is call the breakdown period. After the server is repaired, he returns immediately and provides service until there are no customers in the system.

Let H_o be the complete period of the ordinary M/G/1 queueing system with server breakdowns. Using the well-known result of Kleinrock [24, p. 213], we obtain the expected length of the complete period for the ordinary M/G/1 queueing system with server breakdowns as

$$E[H_o] = \frac{\mu_S(1+\alpha\mu_R)}{1-\rho(1+\alpha\mu_R)}. \quad (\rho = \lambda\mu_S) \tag{4}$$

3.2.1. Expected length of the turned-off period

We know that the turned-off period I_N terminates when the Nth customer arrives in system. Since the length of times between two successive arrivals are independently, identically and exponentially distributed with mean $1/\lambda$, thus the expected length of the turned-off period, $E[I_N]$, for the N policy M/G/1 queueing system with server breakdowns and general startup times is given by

$$E[I_N] = \frac{N}{\lambda}. \tag{5}$$

3.2.2. Expected length of the complete startup period

Let V_N represent the complete startup period for the N policy M/G/1 queueing system with server breakdowns and general startup times. Since the complete startup period is the sum of the complete period and the startup period which implies $V_N = H_N + U_N$, where H_N and U_N denote the complete period and the startup period, respectively. Let $\bar{f}_{V_N}(\cdot)$ be the LST of the distribution of the complete startup period of the N policy M/G/1 queueing system with server breakdowns.

The following notations are used.

- [$F_{V_N}(\cdot)$ –] distribution function of the complete startup period V_N of the N policy M/G/1 queueing system with server breakdowns and general startup times;
- [$\bar{f}_U(\cdot)$ –] the LST of startup time;
- [$F_{H_o}(\cdot)$ –] distribution function of the complete period H_o of the ordinary M/G/1 queueing system with server breakdowns;
- [$F_{H_o}^{(N+n)}(\cdot)$ –] $(N+n)$-fold convolution of $F_{H_o}(\cdot)$.

By conditioning on the length of the startup time U and the number of arrivals during U, we obtain (see [25, p. 277])

$$\begin{aligned} F_{V_N}(x) &= \int_0^x \sum_{n=0}^{\infty} Pr \quad \{\text{given any startup time} = t,\ \text{complete startup period generated by} \\ &\qquad N \text{ customers plus } n \text{ arrivals in the complete period } H_o \text{ during } t \leqslant x-t\}\, dU(t) \\ &= \int_0^x \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} F_{H_o}^{(N+n)}(x-t)\, dU(t). \end{aligned} \tag{6}$$

Taking the LST of both sides of (6) yields

$$\bar{f}_{V_N}(s) = \int_0^{\infty}\int_0^x \sum_{n=0}^{\infty} e^{-sx} \frac{e^{-\lambda t}(\lambda t)^n}{n!} F_{H_o}^{(N+n)}(x-t)\, dU(t)\, dx. \tag{7}$$

Changing the order of integration of (7), it finally gets

$$
\begin{aligned}
\bar{f}_{V_N}(s) &= \int_0^\infty e^{-\lambda t}\,dU(t)\left[\sum_{n=0}^\infty \frac{(\lambda t)^n}{n!}\left[\int_t^\infty e^{-sx}F_{H_o}^{(N+n)}(x-t)\right]dx\right] \\
&= \int_0^\infty e^{-\lambda t}\,dU(t)\left[\sum_{n=0}^\infty \frac{(\lambda t)^n}{n!}e^{-st}[\bar{f}_{H_o}(s)]^{N+n}\right] \\
&= [\bar{f}_{H_o}(s)]^N\int_0^\infty e^{-(s+\lambda)t}\sum_{n=0}^\infty \frac{[\lambda\bar{f}_{H_o}(s)t]^n}{n!}\,dU(t) \\
&= [\bar{f}_{H_o}(s)]^N\int_0^\infty e^{-[s+\lambda-\lambda\bar{f}_{H_o}(s)]t}\,dU(t) \\
&= [\bar{f}_{H_o}(s)]^N\bar{f}_U[s+\lambda-\lambda\bar{f}_{H_o}(s)].
\end{aligned}
\tag{8}
$$

Differentiating (8) with respect to s, we obtain the expected length of the complete startup period as follows:

$$E[V_N] = (N+\lambda\mu_U)E[H_o] + \mu_U = \frac{(N+\lambda\mu_U)\mu_S(1+\alpha\mu_R)}{1-\rho(1+\alpha\mu_R)} + \mu_U. \tag{9}$$

3.2.3. Expected length of the busy and breakdown periods

The expected length of the complete period and the expected length of the startup period are denoted by $E[H_N]$ and $E[U_N]$, respectively. Recall that $V_N = H_N + U_N$ which implies $E[V_N] = E[H_N] + E[U_N]$. Hence from (9) and (4), we obtain

$$E[H_N] = (N+\lambda\mu_U)E[H_o] = \frac{(N+\lambda\mu_U)\mu_S(1+\alpha\mu_R)}{1-\rho(1+\alpha\mu_R)}, \tag{10}$$

$$E[U_N] = \mu_U. \tag{11}$$

Let $E[B_N]$ and $E[D_N]$ be the expected length of the busy period and the expected length of the breakdown period, respectively. Recall that the complete period is the sum of the busy period and the breakdown period which implies $E[H_N] = E[B_N] + E[D_N]$. Hence from (10) we have

$$E[B_N] = \frac{(N+\lambda\mu_U)\mu_S}{1-\rho(1+\alpha\mu_R)}, \tag{12}$$

$$E[D_N] = \frac{(N+\lambda\mu_U)\alpha\mu_S\mu_R}{1-\rho(1+\alpha\mu_R)}. \tag{13}$$

3.3. Expected length of the busy cycle

The busy cycle for the N policy M/G/1 queueing system with server breakdowns and general startup times, denoted by C_N, is the length of time from the beginning of the last turned-off period to the beginning of the next turned-off period. Since the busy cycle is the sum of the turned-off period (I_N), the startup period (U_N), the busy period (B_N), and the breakdown period (D_N), we get

$$E[C_N] = E[I_N] + E[U_N] + E[B_N] + E[D_N] = E[I_N] + E[V_N]. \tag{14}$$

From (5) and (9), we obtain

$$E[C_N] = \frac{N+\lambda\mu_U}{\lambda[1-\rho(1+\alpha\mu_R)]}. \tag{15}$$

3.4. Probability that the server is turned-off, startup, busy and broken down

In steady-state, let

P_{I_N} = probability that the server is turned-off;
P_{U_N} = probability that the server is startup;

P_{B_N} = probability that the server is busy;
P_{D_N} = probability that the server is broken down.

We obtain

$$P_{I_N} = \frac{E[I_N]}{E[C_N]}, \tag{16}$$

$$P_{U_N} = \frac{E[U_N]}{E[C_N]}, \tag{17}$$

$$P_{B_N} = \frac{E[B_N]}{E[C_N]}, \tag{18}$$

$$P_{D_N} = \frac{E[D_N]}{E[C_N]}. \tag{19}$$

Substituting $E[I_N]$ in (5), $E[U_N]$ in (11), $E[B_N]$ in (12), $E[D_N]$ in (13), and $E[C_N]$ in (15) into relations (16)–(19) yields the probability that the server is turned-off, startup, busy and broken down in the following:

$$P_{I_N} = \frac{N(1-\rho_H)}{N+\rho_U}, \tag{20}$$

$$P_{U_N} = \frac{\rho_U(1-\rho_H)}{N+\rho_U}, \tag{21}$$

$$P_{B_N} = \rho, \tag{22}$$

$$P_{D_N} = \alpha\rho\mu_R. \tag{23}$$

We prove from (22) that the probability that the server is busy in the steady-state is equal to ρ.

4. The optimal N policy

We develop an expected cost function per unit time for the N policy M/G/1 queueing system with server startup and breakdowns in which N is a decision variable. Our objective is to determine the optimum value of the control parameter N, say N^*, so as to minimize this function. We define:

C_h = holding cost per unit time for each customer present in the system;
C_s = setup cost for per busy cycle;
C_i = cost per unit time for keeping the server off;
C_{sp} = startup cost per unit time for the preparatory work of the server before starting the service;
C_b = cost per unit time for keeping the server on and in operation;
C_d = breakdown cost per unit time for a failed server.

Utilizing the definition of each cost listed above, the expected cost function per unit time per customer is given by

$$F_o(N) = C_h L_N + C_s \frac{1}{E[C_N]} + C_i \frac{E[I_N]}{E[C_N]} + C_{sp} \frac{E[U_N]}{E[C_N]} + C_b \frac{E[B_N]}{E[C_N]} + C_d \frac{E[D_N]}{E[C_N]}, \tag{24}$$

where L_N is given in (3). We note that $\rho_H + \frac{\lambda^2 E[H^2]}{2(1-\rho_H)}$, $\frac{E[B_N]}{E[C_N]}$ and $\frac{E[D_N]}{E[C_N]}$ do not involve the decision variable N. Omitting these cost terms are not a function of the the decision variable N. The optimization problem in (24) is equivalent to minimize the following equation:

$$F(N) = \frac{1}{N+\rho_U}\left\{\frac{C_h}{2}N^2 - \left[C_h\left(\frac{1}{2}-\rho_U\right) - C_i(1-\rho_H)\right]N + C_h\frac{\lambda^2 E[U^2]}{2} + (C_s\lambda + C_{sp}\rho_U)(1-\rho_H)\right\} \tag{25}$$

Differentiating $F(N)$ with respect to N, we get

$$\frac{dF(N)}{dN} = \frac{C_h}{2} - \frac{C_h}{2(N+\rho_U)^2}\left\{\rho_U + \lambda^2\sigma_U^2 + \frac{2[C_s\lambda + (C_{sp} - C_i)\rho_U]}{C_h}(1-\rho_H)\right\}$$

Setting $dF(N)/dN = 0$ yields

$$N^* = -\rho_U + \sqrt{\rho_U + \lambda^2\sigma_U^2 + \frac{2[C_s\lambda + (C_{sp} - C_i)\rho_U]}{C_h}(1-\rho_H)}, \tag{26}$$

where

$$\rho_U = \lambda\mu_U = \frac{\lambda}{\gamma}\ \left(\mu_U = \frac{1}{\gamma}\right),$$

and

$$\rho_H = \lambda E[H] = \lambda\mu_S(1+\alpha\mu_R) = \frac{\lambda}{\mu}\left(1+\frac{\alpha}{\beta}\right)\left(\mu_S = \frac{1}{\mu}, \mu_R = \frac{1}{\beta}\right).$$

Differentiating $F(N)$ with respect to N twice and using (26), we obtain

$$\frac{d^2F(N)}{dN^2} = C_h\left\{\rho_U + \lambda^2\sigma_U^2 + \frac{2[C_s\lambda + (C_{sp} - C_i)\rho_U]}{C_h}(1-\rho_H)\right\}^{-1/2} > 0, \quad (\rho_H < 1).$$

Thus N^* is the unique minimizer of $F(N)$. If N^* is not an integer, the best positive integer value of N is one of the integers surrounding N^*.

5. Sensitivity analysis

A system analyst often concerns with how the system performance measures can be affected by the changes of the input parameters in the investigated queueing service model. Sensitivity investigation on the queueing model with critical input parameters may provide some answers to this question. In the following, we conduct some sensitivity investigations on the optimal value N^* based on changes in the values of the system parameters $\lambda, \mu, \alpha, \beta, \gamma$ and cost parameters C_h, C_s, C_i, C_{sp}. From (26), we perform some algebraic manipulations with respect to system parameters $\lambda, \mu, \alpha, \beta, \gamma$. Differentiating N^* with respect to λ, we obtain

$$\frac{\partial N^*}{\partial\lambda} = -\mu_U + \frac{(\mu_U + \theta_1)^{1/2}}{2\sqrt{\lambda}}, \quad \text{if } \sigma_U^2 - \theta_1 E[H] = 0, \tag{27}$$

$$\frac{\partial N^*}{\partial\lambda} = -\mu_U + \frac{\mu_U + 2\lambda\sigma_U^2 + \theta_1(1-2\rho_H)}{2\sqrt{\rho_U + \lambda^2\sigma_U^2 + \lambda\theta_1(1-\rho_H)}}, \quad \text{if } \sigma_U^2 - \theta_1 E[H] \neq 0, \tag{28}$$

where

$$\theta_1 = \frac{2[C_s + (C_{sp} - C_i)\mu_U]}{C_h}.$$

Setting (27) and (28) to zero, and then solving for λ, we find

$$\lambda = \frac{\mu_U + \theta_1}{4\mu_U^2}, \quad \text{if } \sigma_U^2 - \theta_1 E[H] = 0, \tag{29}$$

$$\lambda = \frac{\mu_U + \theta_1}{2(\sigma_U^2 - \theta_1 E[H])}\left[-1 + \frac{\mu_U}{\sqrt{\mu_U^2 - \sigma_U^2 + \theta_1 E[H]}}\right], \quad \text{if } \sigma_U^2 - \theta_1 E[H] \neq 0. \tag{30}$$

Note that in (30) the conditions of $\mu_U^2 - \sigma_U^2 + \theta_1 E[H] > 0$ are required. Differentiating (27) and (28) with respect to λ again and substituting (29) and (30) into the resulting differentiation from (27) and (28), respectively, we have

$$\frac{\partial^2 N^*}{\partial \lambda^2} = -\frac{\mu_U^2(\mu_U + \theta_1)}{2} < 0, \quad \text{if } \sigma_U^2 - \theta_1 E[H] = 0, \tag{31}$$

$$\frac{\partial^2 N^*}{\partial \lambda^2} = -\frac{2(\mu_U^2 - \sigma_U^2 + \theta_1 E[H])^{3/2}}{\mu_U + \theta_1} < 0, \quad \text{if } \sigma_U^2 - \theta_1 E[H] \neq 0. \tag{32}$$

The above results show that the graph of N^* is concave downward with respect to λ, which attains its maximum value under two different parameter settings satisfying (29) and (30), respectively. Differentiating N^* with respect to μ yields

$$\frac{\partial N^*}{\partial \mu} = \frac{\rho\theta_1\rho_H}{2\sqrt{\rho_U + \lambda^2\sigma_U^2 + \lambda\theta_1(1-\rho_H)}} > 0, \tag{33}$$

where $\rho = \lambda\mu_S = \lambda/\mu$. Thus, N^* is increasing in μ. Similarly, differentiating N^* with respect to α and β, respectively, we obtain

$$\frac{\partial N^*}{\partial \alpha} = \frac{-\lambda\rho\theta_1}{2\beta\sqrt{\rho_U + \lambda^2\sigma_U^2 + \lambda\theta_1(1-\rho_H)}} < 0, \tag{34}$$

$$\frac{\partial N^*}{\partial \beta} = \frac{\lambda\rho\alpha\theta_1}{2\beta^2\sqrt{\rho_U + \lambda^2\sigma_U^2 + \lambda\theta_1(1-\rho_H)}} > 0. \tag{35}$$

From (34) and (35), we see that N^* is decreasing in α and N^* is increasing in β. If the startup time distribution is given, we can easily see how γ affects N^* because σ_U^2 is a function of the parameter γ. For example, suppose the startup time distribution obeys the Erlang-k ($k > 1$) stage distribution with mean μ_U ($=1/\gamma$). Substituting $\sigma_U^2 = \mu_U^2/k$ into (26) and then differentiating N^* with respect to μ_U, we get

$$\frac{\partial N^*}{\partial \mu_U} = -\lambda + \frac{\frac{2\lambda^2}{k}\mu_U + \lambda\theta_2}{\sqrt{\frac{2\lambda^2}{k}\mu_U^2 + \lambda\mu_U\theta_2 + \lambda\theta_3}}, \tag{36}$$

where

$$\theta_2 = 1 + \frac{2(C_{sp} - C_i)}{C_h}(1-\rho_H) \text{ and } \theta_3 = \frac{2C_s}{C_h}(1-\rho_H).$$

Two situations are considered while investigating the behavior of $\partial N^*/\partial \mu_U$:

Case (i): If $\theta_2^2 - 4\lambda\theta_3 > 0$ and setting $\partial N^*/\partial \mu_U = 0$, then we obtain

$$\mu_U = \frac{k}{2\lambda}\left(-\theta_2 + \sqrt{\frac{k\theta_2^2 - 4\lambda\theta_3}{k-1}}\right), \tag{37}$$

which is a unique solution. Differentiating $\partial N^*/\partial \mu_U$ with respect to μ_U again and using (37), we finally get

$$\frac{\partial^2 N^*}{\partial \mu_U^2} = -\frac{\frac{\lambda^2}{k}(k\theta_2^2 - 4\lambda\theta_3)}{4\left[\frac{\lambda^2}{k}\mu_U^2 + \lambda\mu_U\theta_2 + \lambda\theta_3\right]^{3/2}} < 0. \tag{38}$$

Hence N^* is a concave downward function of μ_U. Since $\mu_U = 1/\gamma$, then it implies that N^* is also a concave downward function of γ. Therefore, from (37), we may obtain

$$\gamma = \frac{2(k-1)\lambda}{k(\theta_2^2 - 4\lambda\theta_3)}\left(\theta_2 + \sqrt{\frac{k\theta_2^2 - 4\lambda\theta_3}{k-1}}\right). \tag{39}$$

Case (ii): If $\theta_2^2 - 4\lambda\theta_3 \leqslant 0$, we can see from (36) that N^* is a decreasing function of μ_U. It also implies that N^* is an increasing function of γ.

In Case (i) and Case (ii) we see how the value N^* is affected by the input parameter γ. On the other hand, it can easily see from (26) that N^* is increasing in C_s, C_{sp} and decreasing in C_i, C_h.

6. Numerical computations

We present some numerical computations to demonstrate the analytical results obtained, and show how to make the decision based on minimizing the cost function (see (25)). Since the cost function is only related to system parameters $\lambda,\mu,\alpha,\beta,\gamma$ in which $\mu_S = 1/\mu$, $\mu_R = 1/\beta$, $\mu_U = 1/\gamma$ and σ_U^2 is a function of γ, then (25) is independent of service time distribution and repair time distribution except for startup time distribution. The sensitivity investigation focuses on the Erlang-2 startup time distribution. First, we fix the following cost parameters $C_s = 1000$, $C_h = 5$, $C_{sp} = 100$, $C_i = 60$ and consider the following five cases.

Case 1: We select $\mu = 0.5,1,1.5,2$, $\alpha = 0.05$, $\beta = 3$, $\gamma = 3$, and vary the values of λ.
Case 2: We select $\lambda = 0.2,0.4,0.6,0.8$, $\alpha = 0.05$, $\beta = 3$, $\gamma = 3$, and vary the values of μ.

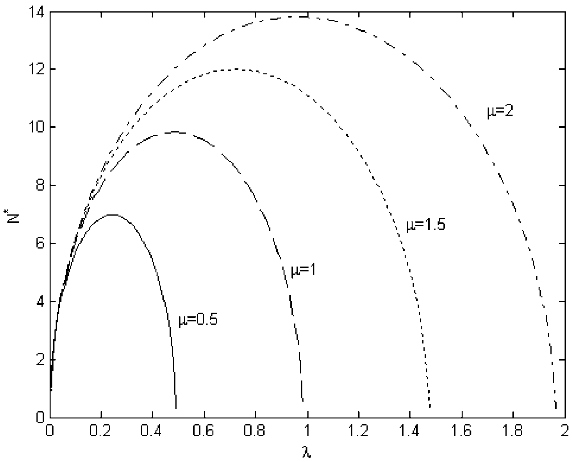

Fig. 1. Plots of (λ,N^*) with $\mu = 0.5$, 1.0, 1.5, 2.0, $\alpha = 0.05$, $\beta = 3$, $\gamma = 3$, $C_s = 1000$, $C_h = 5$, $C_{sp} = 100$, $C_i = 60$.

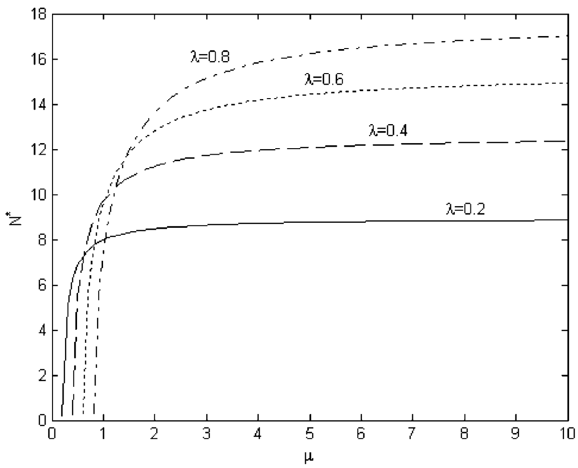

Fig. 2. Plots of (μ,N^*) with $\lambda = 0.2$, 0.4, 0.6, 0.8, $\alpha = 0.05$, $\beta = 3$, $\gamma = 3$, $C_s = 1000$, $C_h = 5$, $C_{sp} = 100$, $C_i = 60$.

Table 1
The optimal N^* and minimum expected $F(N^*)$ with various (λ,μ)

	$\alpha = 0.05,\ \beta = 3,\ \gamma = 3,\ C_s = 1000,\ C_h = 5,\ C_{sp} = 100,\ C_i = 60$							
(λ,μ)	(0.3, 0.5)	(0.3, 1.0)	(0.3, 1.5)	(0.3, 2.0)	(0.2, 1.0)	(0.4, 1.0)	(0.6, 1.0)	(0.8, 1.0)
N^*	7	9	10	10	8	10	10	8
$F(N^*)$	55.3856	85.1964	94.5549	99.1349	85.5034	82.2018	69.7010	47.7634

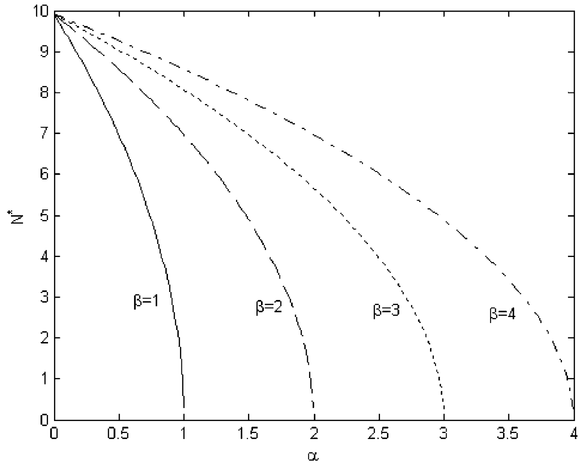

Fig. 3. Plots of (α,N^*) with $\lambda = 0.5,\ \mu = 1,\ \beta = 1,2,3,4\ \gamma = 3,\ C_s = 1000,\ C_h = 5,\ C_{sp} = 100,\ C_i = 60$.

We observe from Fig. 1 that (i) the local maximum value of N^* is moving from left to right as μ increases; and (ii) as λ is fixed, N^* is getting larger as μ increases. From Fig. 2, we see that (i) N^* increases in μ; (ii) if μ is small enough, N^* increases quickly; (iii) if μ is large and $\rho = \lambda/\mu$ is small enough, N^* is insensitive; and (iv) if μ is fixed and large enough, N^* increases in λ. Numerical results of Case 1 and Case 2 are provided in Table 1.

Case 3: We select $\lambda = 0.5,\ \mu = 1,\ \beta = 1,2,3,4,\ \gamma = 3$ and vary the values of α.
Case 4: We select $\lambda = 0.5,\ \mu = 1,\ \alpha = 0.4,0.8,1.2,\ 1.6,\ \gamma = 3$ and vary the values of β.

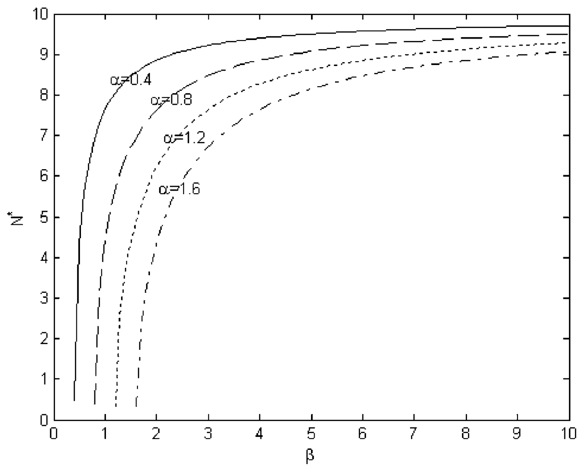

Fig. 4. Plots of (β,N^*) with $\lambda = 0.5,\ \mu = 1,\ \alpha = 0.4,0.8,1.2,1.6,\ \gamma = 3,\ C_s = 1000,\ C_h = 5,\ C_{sp} = 100,\ C_i = 60$.

Table 2
The optimal N^* and minimum expected $F(N^*)$ with various (α,β)

	$\lambda = 0.5,\ \mu = 1,\ \gamma = 3,\ C_s = 1000,\ C_h = 5,\ C_{sp} = 100,\ C_i = 60$							
(α,β)	(0.5, 1.0)	(0.5, 2.0)	(0.5, 3.0)	(0.5, 4.0)	(0.4, 2.0)	(0.8, 2.0)	(1.2, 2.0)	(1.6, 2.0)
N^*	7	9	9	9	9	8	6	4
$F(N^*)$	48.6806	63.8626	68.4962	70.8162	66.7410	54.7399	41.5367	26.2598

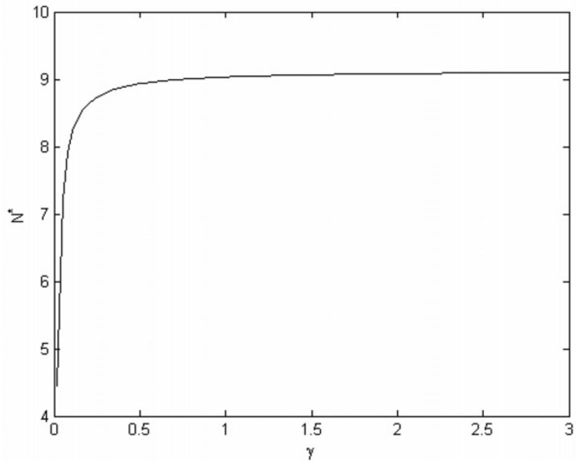

Fig. 5. Plots of (γ,N^*) with $\lambda = 0.3,\ \mu = 1,\ \alpha = 0.05,\ \beta = 3,\ C_s = 1000,\ C_h = 5,\ C_{sp} = 100,\ C_i = 60$.

We observe from Fig. 3 that (i) N^* decreases in α. As α is fixed, the larger β has larger N^*; (ii) N^* has an upper bound as α closes to zero; and (iii) N^* is not insensitive to α. It can easily observe from Fig. 4 that (i) N^* increases in β but N^* is insensitive to β as β is large; and (ii) as β is fixed, the larger α has smaller N^*. Numerical results of Case 3 and Case 4 are listed in Table 2.

Case 5: We select $\lambda = 0.3,\ \mu = 1,\ \alpha = 0.05,\ \beta = 3$ and vary the values of γ.

Fig. 5 indicates that N^* increases in γ if $\theta_2^2 - 4\lambda\theta_3 \leqslant 0$. However, for another set of cost parameters $C_s = 500,\ C_h = 5,\ C_{sp} = 100,\ C_i = 40$. Parameters satisfying $\theta_2^2 - 4\lambda\theta_3 > 0$, N^* has a unique maximum value

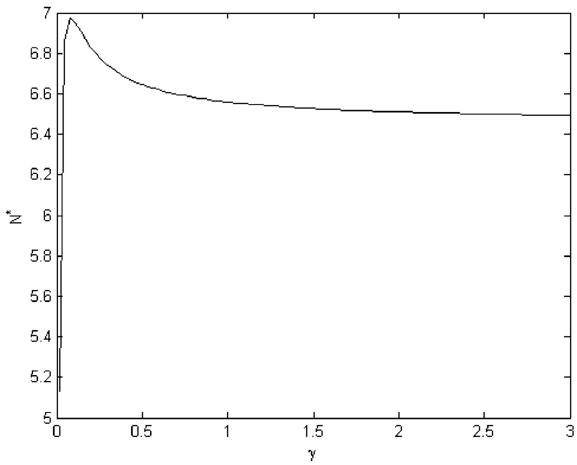

Fig. 6. Plots of (γ,N^*) with $\lambda = 0.3,\ \mu = 1,\ \alpha = 0.05,\ \beta = 3,\ C_s = 500,\ C_h = 5,\ C_{sp} = 100,\ C_i = 40$.

Table 3
The optimal N^* and minimum expected $F(N^*)$ with various γ

	$\lambda = 0.3,\ \mu = 1,\ \alpha = 0.05,\ \beta = 3,\ c_s = 1000,\ c_h = 5,\ c_{sp} = 100,\ c_i = 60$							
γ	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2
N^*	9	9	9	9	9	9	9	9
$F(N^*)$	85.4241	85.2503	85.2181	85.2068	85.2016	85.1987	85.1970	85.1959

Table 4
The optimal N^* and minimum expected $F(N^*)$ with various (c_s,c_h)

	$\lambda = 0.3,\ \mu = 1,\ \alpha = 0.05,\ \beta = 3,\ \gamma = 3,\ c_{sp} = 100,\ c_i = 60$							
(c_s,c_h)	(1000, 5)	(1000, 10)	(1000, 15)	(1000, 20)	(400, 10)	(600, 10)	(800, 10)	(900, 10)
N^*	9	6	5	5	4	5	6	6
$F(N^*)$	85.1964	101.9221	114.0319	124.3333	78.3476	87.3775	95.0861	98.5041

Table 5
The optimal N^* and minimum expected $F(N^*)$ with various (c_{sp},c_i)

	$\lambda = 0.3,\ \mu = 1,\ \alpha = 0.05,\ \beta = 3,\ \gamma = 3,\ c_s = 1000,\ c_h = 5$							
(c_{sp},c_i)	(80, 20)	(80, 30)	(80, 40)	(80, 50)	(35, 25)	(45, 25)	(55, 25)	(65, 25)
N^*	9	9	9	9	9	9	9	9
$F(N^*)$	57.5492	64.4228	71.2964	78.1701	60.6423	60.7187	60.7951	60.8714

at $\gamma = \frac{2(k-1)\lambda}{k(\theta_2^2-4\lambda\theta_3)}\left(\theta_2 + \sqrt{\frac{k\theta_2^2-4\lambda\theta_3}{k-1}}\right)$ (see Fig. 6). Figs. 5 and 6 show that N^* may be too insensitive to changes in γ as γ is greater than 0.4. The numerical results are presented in Table 3.

To see how N^* changes when cost parameter changes, we select $\lambda = 0.3$, $\mu = 1$, $\alpha = 0.05$, $\beta = 3$, $\gamma = 3$, choose $C_{sp} = 100$, $C_i = 60$, and vary the specified values of (C_s,C_h). Table 4 shows that N^* increases in C_s and decreases in C_h. On the other hand, we select $C_s = 1000$, $C_h = 5$ and change the specified values of (C_{sp},C_i). Table 5 reveals that N^* is insensitive to (C_{sp},C_i).

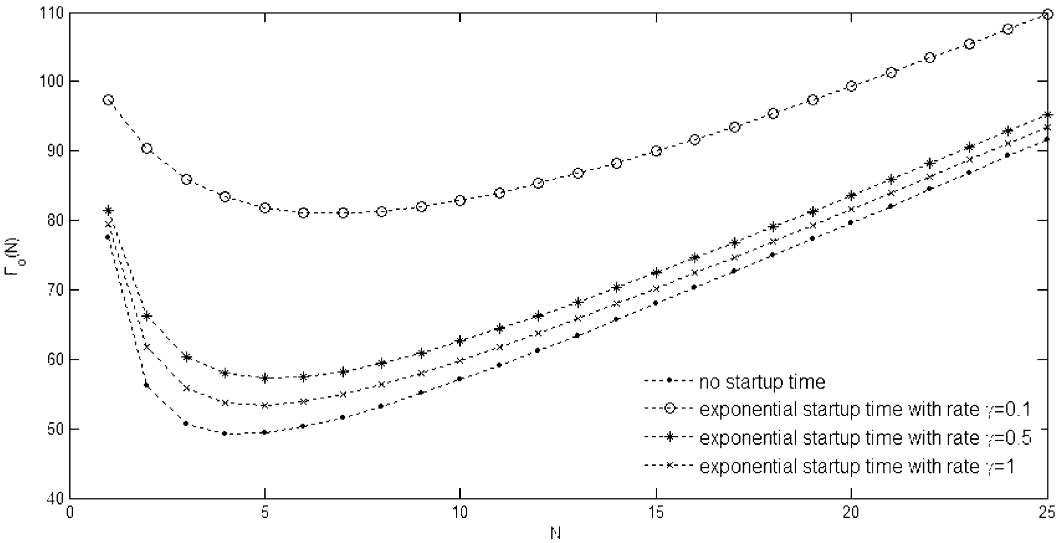

Fig. 7. The total expected cost $F_o(N)$ for different values of γ and N.

Finally, we make comparisons between our model and existing literature (see Pearn et al. [22]). According to the parameters setting by [22], we perform a numerical experiment based on $\lambda = 0.4$, $\mu_S = 1$, $\sigma_S = 1$, $\alpha = 0.05$, $\mu_R = 0.2$, $\sigma_R = 1$, $C_h = 5$, $C_b = 50$, $C_i = 10$, $C_d = 100$ and $C_s = 200$. In addition, we fix startup cost $C_{sp} = 90$ and vary the parameter values (γ) of exponential startup distribution from 0.1 to 1 and N from 1 to 25. Fig. 7 shows that our model approaches to that by [22] as γ tends to large enough (μ_U tends to small enough).

7. Conclusion

In this paper,we considered the optimal control N policy for the M/G/1 queueing system with server breakdowns and general distributed startup times. We developed the theoretical results for system performance measures, such as the expected number of customers in system, the expected length of the turned-off, complete startup, busy, and breakdown periods, and the expected length of the busy cycle. In the N policy M/G/1 queueing system with general service times and startup times, we proved that the probability that the server is busy in the steady-state is equal to the traffic intensity ρ. We constructed a cost model to determine the optimal threshold N so as to minimize an expected cost function. We also provided sensitivity analysis to discuss how the system performance measures can be affected by the changes of the input parameters (or cost parameters) in the investigated queueing service model.

References

[1] M. Yadin, P. Naor, Queueing systems with a removable service station, Oper. Res. Quart. 14 (1963) 393–405.
[2] D.P. Heyman, Optimal operating policies for M/G/1 queueing system, Oper. Res. 16 (1968) 362–382.
[3] C.E. Bell, Characterization and computation of optimal policies for operating an M/G/1 queueing system with removable server, Oper. Res. 19 (1971) 208–218.
[4] H.C. Tijms, Stochastic Modelling and Analysis, Wiley, New York, 1986.
[5] K.-H. Wang, J.-C. Ke, A recursive method to the optimal control of an M/G/1 queueing system with finite capacity and infinite capacity, Appl. Math. Modell. 24 (2000) 899–914.
[6] K.-H. Wang, H.-M. Huang, Optimal control of an M/E_k/1 queueing system with a removable service station, Oper. Res. Soc. 46 (1995) 1014–1022.
[7] K.-H. Wang, K.-L. Yen, Optimal control of an M/H_k/1 queueing system with a removable server, Math. Methods Oper. Res. 57 (2002) 255–262.
[8] K.R. Baker, A note on operating policies for the queue M/M/1 with exponential startup, INFOR 11, 1973, pp. 71–72.
[9] A. Borthakur, J. Medhi, R. Gohain, Poisson input queueing systems with startup time and under control operating policy, Comput. Oper. Res. 14 (1987) 33–40.
[10] J. Medhi, J.G.C. Templeton, A poisson input queue under N-policy and with a general start up time, Comput. Oper. Res. 19 (1992) 35–41.
[11] H. Takagi, A M/G/1/K queues with N-policy and setup times, Queueing Syst. 14 (1993) 79–98.
[12] H.W. Lee, J.O. Park, Optimal strategy in N-policy production system with early set-up, J. Oper. Res. Soc. 48 (1997) 306–313.
[13] Reddy G.V. Krishna, R. Nadarajan, R. Arumuganathan, Analysis of a bulk queue with N-policy multiple vacations and setup times, Comput. Oper. Res. 25 (1998) 957–967.
[14] S. Hur, S.J. Paik, The effect of different arrival rates on the N-policy of M/G/1 with server setup, Appl. Math. Modell. 23 (1999) 289–299.
[15] K.-H. Wang, Optimal operation of a Markovian queueing system with a removable and non-reliable server, Microelectron. Reliab. 35 (1995) 1131–1136.
[16] K.-H. Wang, Optimal control of an M/E_k/1 queueing system with removable service station subject to breakdowns, J. Oper. Res. Soc. 48 (1997) 936–942.
[17] K.-H. Wang, K.-W. Chang, B.D. Sivazlian, Optimal control of a removable and non-reliable server in an infinite and a finite M/H_2/1 queueing system, Appl. Math. Modell. 23 (1999) 651–666.
[18] K.-H. Wang, H.-T. Kao, G. Chen, Optimal management of a removable and non-reliable server in an infinite and a finite M/H_k/1 queueing system, Quality Technol. Quant. Manage. 1 (2) (2004) 325–339.
[19] K.-H. Wang, J.-C. Ke, Control policies of an M/G/1 queueing system with a removable and non-reliable server, Int. Trans. Oper. Res. 9 (2002) 195–212.
[20] J.-C. Ke, The optimal control of an M/G/1 queueing system with server vacations, startup and breakdowns, Comput. Indust. Eng. 44 (2003) 567–579.
[21] J.-C. Ke, W.L. Pearn, Optimal management policy for heterogeneous arrival queueing systems with server breakdowns and vacations, Quality Technol. Quant. Manage. 1 (1) (2004) 149–162.

[22] W.L. Pearn, J.-C. Ke, Y.C. Chang, Sensitivity analysis of optimal management policy for a queueing system with a removable and non-reliable server, Comput. Indust. Eng. 46 (2004) 87–99.
[23] K.-H. Wang, Optimal control of a removable and non-reliable server in an M/M/1 queueing system with exponential startup time, Math. Methods Oper. Res. 58 (2003) 29–39.
[24] L. Kleinrock (Ed.), Queueing Systems, vol. I, Wiley, New York, 1976.
[25] D. Gross, C.M. Harris, Fundamentals of Queueing Theory, Wiley, New York, 1985.