Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Computers & Operations Research 31 (2004) 1001–1015

computers & operations research

www.elsevier.com/locate/dsw

# Optimal management of the $N$-policy $M/E_k/1$ queuing system with a removable service station: a sensitivity investigation

W.L. Pearn*, Y.C. Chang

*Department of Industrial Engineering and Management, 1001 Ta Hsueh Road, Hsin Chu 30050, National Chiao Tung University, Taiwan, ROC*

**Abstract**

This paper deals with optimal management problem of the $N$-policy $M/E_k/1$ queuing system with a removable service station under steady-state condition. The server is in a controllable position that the manager can turn the single server on at any arrival epoch, or off at any service completion. Arrival time and service time of the customers are assumed to follow the negative exponential distribution and the Erlang $k$ type distribution, respectively. In this paper, we consider a practical application of such model. A cost formula is established to determine the optimal management policy of the removable service station to minimize the total expected cost per customer per unit time. We apply the analytic solution of the queuing model and use an efficient Matlab computer program to calculate the optimal value of $N$ and some system performance measures. Analytical results for sensitivity analysis are derived. We provide extensive numerical computation for illustration purpose, and demonstrate how the model could be used in real applications.

*Keywords:* Analytical results; Queues; Management policy; Sensitivity analysis

## 1. Introduction

This paper considers an $M/E_k/1$ queuing system with a removable service station. The term 're-movable server' states the operation policy of the system allows one to turn on and turn off the server, depending on the number of customers in the system. The Erlang distribution, denoted by $E_k$ is a special case of the gamma distribution, is named after A.K. Erlang who pioneered queuing systems theory for its application to congestion in telephone networks. The Erlang was an early, but successful distribution used in establishing a queuing model without assuming an exponential distribution for service-times, but still keeping most of the $M/M/1$ properties mathematical tractable.

* Corresponding author. Tel.: +886-35-714261; fax: +886-35-722392.
*E-mail address:* roller@cc.nctu.edu.tw (W.L. Pearn).

0305-0548/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0305-0548(03)00048-0

The server is removable and applies the $N$-policy. That is, the server operation starts only when $N(N \geqslant 1)$ customers have accumulated, and is shut down (turned off) when no customer is present. After the server is turned off, the server may not operate until $N$ customers are present in the system.

The management policy that the decision-maker can turn a single service station on at the customers' arrival epochs or off at service completion epochs is investigated. The queuing problem with removable service station has been extensively studied in the literature. A pioneer work in this field is Yadin and Naor [1], who first introduced the concept of an $N$-policy, which turns the server on when the number of customers in the system reaches a certain number, $N(N \geqslant 1)$, and turns the server off when there is no customer in the system. Several types of queuing models with single-removable server, have been investigated by Bell [2,3], Heyman [4], Kimura [5], Sobel [6], Teghem [7] and many others, under various assumptions on the interarrival and service time distributions. The queuing model undertaken in this paper, generalize the previous results on the controllable $M/M/1$ queuing system by Sivazlian and Stanfel [8], the ordinary $M/M/1$ queuing system by Gross and Harris [9], and the ordinary $M/E_k/1$ queuing system by Gross and Harris [9].

We construct the total expected cost function per unit time, where the cost element consists of (i) a holding cost for waiting customers; (ii) an operating cost for operating the service station; (iii) a start-up cost and a removable cost for activating and removing the service station, respectively, and (iv) a cost for performing the auxiliary task by the service station. Finally, a cost formula is developed. One is then interested in determining the optimal value of the decision variable $N$ to minimize the long-run expected cost per unit time. The primary objectives of this paper are (i) to use an efficient Matlab computer program to calculate the optimal value of $N$ and other critical system performance measures, and (ii) to derive the analytical results on the sensitivity analysis. We carry out extensive computational experiments to illustrate the analytical sensitivity results, and present an application example demonstrating how the computer program such as Matlab can be applied to calculate the system performance measures, the optimum value of the management parameter $N$, and its minimum expected cost under various system parameter values of consideration, while maintaining the minimal service quality. The result is useful to the managers for making reliable decisions in managing their service systems.

## 2. The $M/E_k/1$ queue with removable service station

We consider the following queuing model formulation (see Wang [10]). A busy cycle of the model consists of an idle period and a busy period. When the system is empty, one busy cycle begins. The server remains in turned-off status until there are $N$ customers in the system. We call this the idle period. The busy period is initiated when the server starts serving the customers waiting in the system. The busy cycle starts with the idle period, and terminates when all the customers are served, the busy cycle may be represented as the sum of the idle period and the busy period. It is assumed that the customers arrive following a Poisson process with parameter $\lambda$, with service times following an Erlang distribution with mean $1/\mu$ and stage parameter $k$. The Erlang type $k$ distribution is made up of $k$ independent and identical exponential stages, each with mean $1/k\mu$.

A customer goes into the first service station (say stag 1), then progresses through the remaining service stations and completes the service at the last service station (say stage $k$) before the next customer enters the first service station. We assume that the server can only serve one customer at a time, and it takes a zero set-up time to restart the service station. Customers arriving at the service station form a single waiting line and are served in the order of their arrivals, that is, in the first-come first-served discipline (FCFS). It is further assumed that the service is independent of the arrival of the customers. If the service station is busy, then a newly arriving customer or waiting customers must wait in the queue until the station is available. Whenever the system is empty the idle period starts. When the server finds at least $N$ customers waiting in the system, the server begins the service immediately until the system becomes empty again.

## 2.1. Steady-state solutions

### 2.1.1. Expected number of customers in the system

We define the expected number of customers in the $M/E_k/1$ queuing system under the $N$-policy as follows: $L_{off} \equiv$ the expected number of customers in the system when the service station is turned off, $L_{on} \equiv$ the expected number of customers in the system when the service station is turned on; $L_N \equiv$ the expected number of customers in the system. From Wang [10] we have the following analytic closed- form expressions, where $\rho = \lambda/\mu$ and $r = \lambda/k\mu$:

$$L_{off} = \frac{(N-1)(1-\rho)}{2}, \tag{1}$$

$$L_{on} = \frac{\rho(N+1-\rho N+r)}{2(1-\rho)}, \tag{2}$$

$$L_N = \frac{N-1}{2} + \frac{\rho(r-\rho+2)}{2(1-\rho)}. \tag{3}$$

### 2.1.2. Long run fraction of time measures

Notations for idle period, the busy period, and the busy cycle are defined as follows. The idle period, the length of time the service station is turned off per cycle, is denoted by $I$. The busy period, the length of time the service station is turned on in operation, and the customers are being served per cycle, is denoted by $B$. The busy cycle, from the beginning of the last idle period to the beginning of the following next idle period, is denoted by $C$. The expected lengths of the idle period, the busy period, and the busy cycle, are denoted by $E[I]$, $E[B]$ and $E[C]$, respectively. The busy cycle is the sum of the idle period and the busy period, $C = I + B$, or $E[C] = E[I] + E[B]$. Using the results stated in Wang [10], we have the long-run fraction of time, for the server is in idle, busy, a busy cycle, respectively, and the number of cycles per unit time are

$$\frac{E[I]}{E[C]} = 1-\rho, \tag{4}$$

$$\frac{E[B]}{E[C]} = \rho, \tag{5}$$

$$E[C] = \frac{N}{\lambda(1-\rho)}, \tag{6}$$

$$\frac{1}{E[C]} = \frac{\lambda(1-\rho)}{N}. \tag{7}$$

Empty probability, that there is no customer in the system and no station is in service (the service station is turned off), is given by

$$P_{00}^{0}(0) = \frac{1-\rho}{N}. \tag{8}$$

Stability conditions for a stable queuing system are given by Eq. (8) with $0 < P_{00}^{0}(0) < 1$. With simple algebraic manipulations, we obtain the following inequality, where $\rho = \lambda/\mu$, which is sufficient for stationary conditions.

$$0 < \rho < 1. \tag{9}$$

For the more general case, $N$-policy $M/G/1$ queue with a removable server under the steady-state condition, we remark that the stationary system performance measures such as the long-run fraction of time for the server idle or busy, and the number of busy cycles per unit time, are identical to those of $N$-policy $M/E_k/1$ queuing system with a removable server (see Wang and Ke [11]), but the results on the expected number of customers in the system are different.

## 3. Optimal management policy

In this section, we develop the total expected cost function per unit time for the $M/E_k/1$ queuing system, in which $N$ is a management decision variable. Following the construction of the cost function, our objective is to determine the optimal value of the management parameter $N$, denoted as $N^*$, to minimize this total expected cost function. We define the following: (1) $C_\text{h} \equiv$ holding cost per unit time for each customer presently in the system. The holding cost can be treated as the penalty cost for delaying service to the customers waiting in the system for service, (2) $C_\text{o} \equiv$ operating cost per unit time for the service station in operation. The operating cost is incurred by the operating service station to provide service for the customers, (3) $C_\text{s} \equiv$ start-up cost per unit time for activating the service station while the service station is turned off (or is removed from the system). The start-up cost is incurred each time the service station starts a new operation when the service station is in turned-off status, (4) $C_\text{r} \equiv$ removable cost per unit time for removing the service station from the service. The removable cost is incurred each time the operating service station is removed from the system, (5) $C_\text{a} \equiv$ cost per unit time for performing an auxiliary task by the service station.

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

$$TC(N) = C_\text{h} L_N + C_\text{o}\frac{E[B]}{E[C]} + (C_\text{s} + C_\text{r})\frac{1}{E[C]} + C_\text{a}\frac{E[I]}{E[C]}. \tag{10}$$

We should note that the second term of Eq. (3) is not a function of the decision variable $N$. Likewise, we note from Eqs. (4)–(5) that, terms $E[B]/E[C]$, and $E[I]/E[C]$ do not involve the decision variable $N$. Omitting those cost terms not a function of the decision variable $N$, the optimization problem in (10) is equivalent to minimizing the following equation:

$$\tilde{T}C(N) = C_h \frac{N-1}{2} + (C_s + C_r)\frac{\lambda(1-\rho)}{N}. \tag{11}$$

Discarding the fixed cost $-(1/2)C_h$ of the first term, Eq. (11) reduces to the following expression, subject to $0 < \rho < 1$, and $N = 1, 2, \ldots$

$$\hat{T}C(N) = C_h \frac{N}{2} + (C_s + C_r)\frac{\lambda(1-\rho)}{N}. \tag{12}$$

### 3.1. Determine the optimal management policy

Since $N$ is a positive integer, $N = 1, 2, \ldots,$ the optimal value $N^*$ minimizing $TC(N)$ can be determined from the following two inequalities,

$$\begin{aligned}&\hat{T}C(N^* - 1) \geqslant \hat{T}C(N^*),\\ &\hat{T}C(N^* + 1) \geqslant \hat{T}C(N^*).\end{aligned} \tag{13}$$

From (12), the necessary conditions for $N^*$ to be optimal reduce to

$$(N^* - 1) \leqslant \frac{2\lambda(C_s + C_r)(1-\rho)}{C_h} \leqslant N^*(N^* + 1). \tag{14}$$

The optimal value $N^*$ may be determined by giving a particular value of $2\lambda$ $(C_s + C_r)(1-\rho)/C_h$. Note that there might be two simultaneous solutions for Eq. (14) which minimize the total expected cost function $TC(N)$. For example, we set a particular value of $2\lambda(C_s + C_r)(1-\rho)/C_h = 30$ in Eq. (14) and solve for $N^*$ to obtain $N^* = 5$ or 6. If $N$ is treated as a continuous variable greater than zero, we present two methods to solve for the optimal of $N$, say $N^*$, and convexity of $TC(N)$ will be proved. Note that the MATLAB computer program we used allows one to plot $TC(N)$ versus $N^*$ to illustrate the convexity property (see Fig. 4).

*Method 1*. Differentiate $TC(N)$ with respect to $N$ and setting the result equal to zero yields

$$\frac{C_h}{2} - (C_s + C_r)\frac{\lambda(1-\rho)}{N^2} = 0.$$

Thus, the optimal value of $N$ is approximately given by

$$N^* = \left(\frac{2\lambda(1-\rho)(C_s + C_r)}{C_h}\right)^{1/2}. \tag{15}$$

Differentiate $TC(N)$ with respect to $N$ twice and then substitute

$$N^* = \left(\frac{2\lambda(1-\rho)(C_s + C_r)}{C_h}\right)^{1/2} \quad \text{to obtain}$$

$$\frac{d^2TC(N^*)}{dN^2} = \sqrt{\frac{C_h^3}{2\lambda(C_s + C_r)(1-\rho)}} > 0, \quad \text{for } \rho < 1, \tag{16}$$

which implies that $TC(N)$ is a concave upward (convex) function and achieves a global minimum when

$$N^* = \left(\frac{2\lambda(1-\rho)(C_s+C_r)}{C_h}\right)^{1/2}. \tag{17}$$

*Method 2*. From (12) we have the following inequality

$$\hat{T}C(N) = C_h\frac{N}{2} + (C_s+C_r)\frac{\lambda(1-\rho)}{N} \geqslant \sqrt{2\lambda C_h(C_s+C_r)(1-\rho)}, \tag{18}$$

which gives a lower bound of $\hat{T}C(N)$ and indicates that $\hat{T}C(N)$ is a concave upward function with lower bound $\sqrt{2\lambda C_h(C_s+C_r)(1-\rho)}$. Equality in (18) holds when

$$C_h\frac{N}{2} = (C_s+C_r)\frac{\lambda(1-\rho)}{N}. \tag{19}$$

With some algebraic manipulations, we obtain

$$N^* \approx \left(\frac{2\lambda(1-\rho)(C_s+C_r)}{C_h}\right)^{1/2}. \tag{20}$$

Note that the expressions of $N^*$ in Eqs. (17) and (20) are the same. If $N^*$ is not an integer, the optimal value $N^*$ may be found the integer closest to the following expression,

$$N^* = \left(\frac{2\lambda(1-\rho)(C_s+C_r)}{C_h}\right)^{1/2} + \varepsilon, \tag{21}$$

where $\varepsilon \in (-1, 1)$ is a constant.

For the more general case, $N$-policy $M/G/1$ queuing system with a removable server, where the service times are assumed to follow the general distributions, with some similar algebraic manipulations it is interesting to note that we would obtain the same expression for the optimal value $N^*$ as stated in Eq. (21).

## 4. Analytical results for sensitivity analysis

A system analyst often concern with how the system performance can be affected by the changes of the input parameters in the recommended queuing service model. Sensitivity investigation on the queuing 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 values of the cost parameters $C_h$, $C_o$, $C_s$, $C_r$, $C_a$ and system parameters $\lambda$, $\mu$, and $k$.

We note that the terms $E[B]/E[C]$, and $E[I]/E[C]$ do not involve the decision variable $N$. Therefore, we may set the relative cost parameters $C_o$ and $C_a$ to be some fixed constants. Further, from Eq. (18), it is easy to see that

$$N^* \propto \sqrt{(C_s+C_r)/C_h}.$$

We perform some algebraic manipulation with respect to system parameters $\lambda$, and $\mu$. By differentiate $N^*$ with respect to $\lambda$, we obtain

$$\frac{\partial N^*}{\partial \lambda} = \frac{(1-2\rho)\sqrt{(C_s + C_r)}}{\sqrt{2\lambda C_h(1-\rho)}}. \tag{22}$$

Set the last equation equal to 0 then solve for $\lambda$, we find $\lambda = \mu/2$ (note that the condition of $\lambda < \mu$ is required). By differentiating $\partial N^*/\partial \lambda$ with respect to $\lambda$ again and substitute $\lambda = \mu/2$, we can easily show that

$$\left.\frac{\partial^2 N^*}{\partial \lambda^2}\right|_{\lambda=\mu/2} = -2\sqrt{\frac{2(C_s + C_r)}{C_h \mu^3}} < 0. \tag{23}$$

The above result implies that $N^*$ is a concave downward function with respect to $\lambda$, which achieves its maximum at $\lambda = \mu/2$. By differentiate $N^*$ with respect to $\mu$, we have

$$\frac{\partial N^*}{\partial \mu} = \frac{\rho^2\sqrt{C_s + C_r}}{\sqrt{2\lambda C_h(1-\rho)}} > 0, \tag{24}$$

for $\lambda < \mu$, $\forall \mu$. Thus, $N^*$ is increasing in $\mu$. It is interesting to note that the stage parameter $k$ do not appear in the expression (21). The result implies that the decision variable $N^*$ is insensitive to the number of stages $k$. To sum up, we have the following analytical results for the sensitivity analysis.

(1) $N^*$ increases in $\lambda$ for $\rho < 1/2$ and decreases in $\lambda$ for $\rho > 1/2$.
(2) $N^*$ increases in $\mu$.
(3) Stage parameter $k$, $C_o$ and $C_a$ do not affect $N^*$.
(4) $N^*$ is proportional to $\sqrt{(C_s + C_r)/C_h}$. In other words, $N^*$ increases in $C_s$ and $C_r$ whereas decreases in $C_h$.

The results show some interest properties of the $M/E_k/1$ queuing system with a removable service station. For low traffic intensity service systems with $\rho < 1/2$ (sparse system): when arrival customers increase, we should raise the threshold $N^*$ to start serving waiting customers. On the other hand, for high traffic intensity service systems with $\rho > 1/2$ (crowded system): when arrival customers increase, we should reduce the threshold $N^*$ to start serving waiting customers to maintain low cost. For the service station, as long as it can serve in a faster rate, the system manager should increase the threshold $N^*$. Stage parameter $k$ do not influence the decision variable $N$. We recall that the definition of the Erlang distribution states that a customer completes the service at the last service station (say stage $k$) before the next customer enters the first service station. So only the total service rate $1/\mu$ is concerned. Operating cost per unit time for the service station in operation and cost per unit time for performing an auxiliary task by the service station may treat as fixed cost to the service system and they would not affect the decision variable $N$. We should note that $N$-policy is used because of expensive start-up and shut down cost per cycle (relative to holding cost), they affect $N^*$ in the following way: for the same cost ratio (cost per cycle relative to holding cost), we would obtain the same value $N^*$. As start-up cost per unit time for activating the service station or removable cost per unit time increase, one should increase the threshold $N^*$ to prevent such set-up and shut down costs. When holding cost per unit time for each customer presently in the system increase, we should decrease the threshold $N^*$ to avoid heavy holding cost.

## 5. Numerical computations

Based on changes in considerable values of the cost parameters $C_h$, $C_o$, $C_s$, $C_r$, $C_a$ and system parameters $\lambda$, $\mu$ and $k$, we now perform a numerical illustration for the analytical results for the sensitivity analysis on the optimum value $N^*$. It should be noted that the terms $E[B]/E[C]$ and $E[I]/E[C]$ do not involve the decision variable $N$. We may set the corresponding cost parameters $C_o$, and $C_a$ to be some fixed constants. Additionally, incremental, rather than accounting costs are considered, since the latter often include such non-incremental elements as overhead. So $C_h$ is set to be \$5, \$10, \$20, \$40, and \$80 to cover various level of, from low to high the holding costs. Eq. (18) suggests that $N^* \propto \sqrt{(C_s + C_r)/C_h}$. We note that $N$ policy is applied to manage the queuing system due to expensive start-up and shut down cost per cycle (relative to holding cost). We may treat $(C_s + C_r)$ as the cost per cycle, without loss of generality, we assume $C_s$ and $C_r$ to be equal since only the sum of them is concerned. Also, the ratio $(C_s + C_r)/C_h$ is set to 40, 80, 160, 320, and 640, to cover five levels of cost relationship (Cases 1–5 and Cases 9–5). The numerical values are obtained by considering the cost parameters as tabulated in Table 1.

We now consider the following experimental design of system parameters for sensitivity analysis on the optimum value $N^*$ based on changes in considerable values of $\lambda$, $\mu$ and $k$. Note that $0 < \rho < 1$ is sufficient for steady-state condition. We calculate the optimal value $N^*$, and the corresponding minimum expected cost $TC(N)$ for the parameters settings listed in Table 2, which cover a wide range of applications dealing with the referred queuing model. A queuing system may be characterized by $\rho = \lambda/\mu$ which represents the traffic intensity. In our investigation, $\rho \in (0.1,\ 0.9)$, and stage parameter $k = 6$ are considered.

Rows 2, 3, 4 list the parameter settings for various $\lambda$. For specified traffic intensity $\rho$ varies from 0.1 to 0.9 (low to high) and three levels of $\mu = 1$, 2, and 3. Solve $\rho = \lambda/\mu$ for $\lambda$ to obtain $\lambda = 0.1(0.05)0.9$, $\lambda = 0.2(0.1)1.8$, and $\lambda = 0.3(0.15)2.7$. Rows 5, 6, 7 list the parameter settings for various $\mu$. We consider traffic intensity $\rho$ varies from 0.86 to 0.1 (high to low) and three levels

Table 1
Cost parameter values considered

| Case | $C_h$ | $C_o$ | $C_a$ | $C_s = C_r$ |
|---|---|---|---|---|
| 1 | 5 | 50 | 10 | 100 |
| 2 | 5 | 50 | 10 | 200 |
| 3 | 5 | 50 | 10 | 400 |
| 4 | 5 | 50 | 10 | 800 |
| 5 | 5 | 50 | 10 | 1600 |
| 6 | 10 | 50 | 10 | 1600 |
| 7 | 20 | 50 | 10 | 1600 |
| 8 | 40 | 50 | 10 | 1600 |
| 9 | 80 | 50 | 10 | 1600 |

Table 2
Parameters settings for various system parameter combinations based on specified $\rho$

| $\lambda$ | $\mu$ | Parameter setting | Description of $\rho$ |
|---|---|---|---|
| — | 1 | $\lambda = 0.1(0.05)0.9$ | $\rho = 0.1(0.05)0.9$ |
| — | 2 | $\lambda = 0.2(0.1)1.8$ | $\rho = 0.1(0.05)0.9$ |
| — | 3 | $\lambda = 0.3(0.15)2.7$ | $\rho = 0.1(0.05)0.9$ |
| 0.4 | — | $\mu = 0.45(0.25)4.45$ | $\rho \in (0.09, 0.89)$ |
| 0.6 | — | $\mu = 0.7(0.35)6.3$ | $\rho \in (0.1, 0.86)$ |
| 0.8 | — | $\mu = 0.9(0.5)8.9$ | $\rho \in (0.09, 0.89)$ |

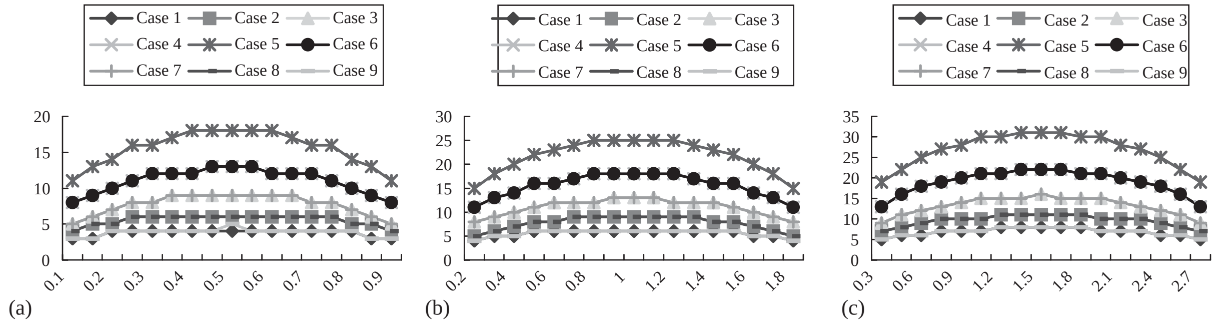

Fig. 1. (a) Plots of $N^*$ versus $\lambda$ for $\mu = 1$, and case 1(1)5 (bottom to top in plot). (b) Plots of $N^*$ versus $\lambda$ for $\mu = 2$, and case 1(1)5 (bottom to top in plot). (c) Plots of $N^*$ versus $\lambda$ for $\mu = 3$, and case 1(1)5 (bottom to top in plot).

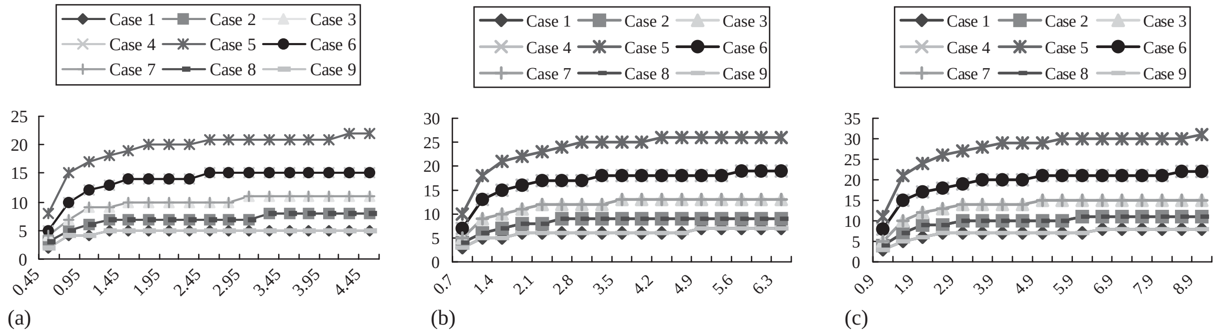

Fig. 2. (a) Plots of $N^*$ versus $\mu$ for $\lambda = 0.4$, and case 1(1)5 (bottom to top in plot). (b) Plots of $N^*$ versus $\mu$ for $\lambda = 0.6$, and case 1(1)5 (bottom to top in plot). (c) Plots of $N^*$ versus $\mu$ for $\lambda = 0.8$, and case 1(1)5 (bottom to top in plot).

of $\lambda = 0.4$, 0.6, and 0.8. Then we solve $\rho = \lambda/\mu$ for $\mu$ to get $\mu = 0.45(0.25)4.45$, $\mu = 0.7(0.35)6.3$, and $\mu = 0.9(0.5)8.9$. Figures are arranged in the following way: Fig. 1(a)–(c) plot the curves of $N^*$ versus $\lambda$ for various cost cases in Table 1 and parameters settings in Table 2. Fig. 2(a)–(c) plot the curves of $N^*$ versus $\mu$ for various cost cases in Table 1 with parameters settings in Table 2.

### *5.1. Interpretation of the results in tables*

The optimal value, $N^*$, and the corresponding minimum expected cost $TC(N^*)$ are displayed in Tables 3 for parameter $\mu = 2$ and $\lambda = 0.2(0.1)1.8$, as the case shown in Table 2 (see row 3). From Table 3 we observe that (i) $N^*$ increases in $\lambda$ for $\rho < 1/2$, and decreases in $\lambda$ for $\rho > 1/2$, (ii) $TC(N^*)$ increases as $\lambda$ increases, (iii) $N^*$ decreases but $TC(N^*)$ increases as $C_h$ increases for fixed values of $C_o$, $C_a$, $C_s$ and $C_r$ (Cases 5–9 in Table 1), (iv) $N^*$ and $TC(N^*)$ both increase as $C_s$ and $C_r$ increase for fixed values of $C_h$, $C_o$ and $C_a$ (Cases 1–5 in Table 1).

The optimal value, $N^*$, and the corresponding minimum expected cost $TC(N^*)$ are displayed in Table 4, for parameter $\lambda = 0.6$, and $\mu = 0.7(0.35)6.3$, as shown in Table 2 (see row 6). From Table 4, we observe that (i) $N^*$ increases as $\mu$ increases, (ii) $TC(N^*)$ decreases as $\mu$ increases, (iii) $N^*$ decreases but $TC(N^*)$ increases as $C_h$ increases for fixed values of $C_o$, $C_a$, $C_s$ and $C_r$ (Cases 5–9 in Table 1). (iv) $N^*$ and $TC(N^*)$ both increase as $C_s$ and $C_r$ increase for fixed values of $C_h$, $C_o$ and $C_a$ (Cases 1–5 in Table 1). From our numerical investigations, the results coincide with our analytical results.

### *5.2. Interpretation of the results in figures*

Fig. 1(a)–(c) reveal that: (i) $N^*$ increases in $\lambda$ for $\rho < 1/2$ and decreases in $\lambda$ for $\rho > 1/2$, (ii) $N^*$ increases in $C_s$ and $C_r$ but decreases in $C_h$. From Fig. 2(a)–(c) we observe that: (i) $N^*$ increases in $\mu$, (ii) $N^*$ increases in $C_s$ and $C_r$ but decreases in $C_h$. We see that Cases 1 and 9; 2 and 8; 3 and 7; 4 and 6 overlap (except for few points in all figures). It is noted that the cost ratio $(C_s + C_r)/C_h$ for these pairs 1 and 9; 2 and 8; 3 and 7; 4 and 6, are 40, 80, 160, 320, respectively. It seems that we would obtain the same value $N^*$ for the same cost ratio $(C_s + C_r)/C_h$.

## 6. An application example

A practical example related to computer communication networks is presented in the following for illustrative purpose. The idea is that the service facility consists of a number of stages, and the customers have to pass through each stage of the service station. The next customer will not be taken for service until the previous customer has completed all the stages. The stages may, in some cases, correspond to how service is actually provided, or they may be purely conceptual. The time taken for service at each stage follows an exponential distribution, and the mean time between any two stages is the same. The distribution of service-time for the whole facility is said to be Erlang-$k$, where $k$ is the number of stages.

Computer communication networks use a variety of flow control policies to achieve high performance (throughput), low delay, and good stability. Here, we model the flow control policy of IBM's System Network Architecture (SNA). SNA routes messages from sources to destinations by way of intermediate nodes, which temporarily buffer the messages. Messages buffers are a scarce resource. The flow control policy regulates the flow of messages between source/destination pairs in an effort to avoid problems such as deadlock and starvation, which could result from poor buffer management. SNA has a window flow control policy, and the key control parameter is the window size, $N$. Under FCFS scheduling, messages arrive at the service center, wait in the pacing box and queue for service

Table 3
The values of $N^*$, and the minimum cost $TC(N^*)$ for $\mu = 2$

| $\lambda$ | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case 1 | 4,31 | 5,37 | 5,42 | 6,46 | 6,50 | 6,54 | 6,57 | 6,60 | 6,63 | 6,66 | 6,68 | 6,70 | 6,73 | 6,75 | 5,78 | 5,82 | 4,91 |
| Case 2 | 5,39 | 6,46 | 7,52 | 8,58 | 8,62 | 9,67 | 9,70 | 9,73 | 9,76 | 9,79 | 9,81 | 9,83 | 8,85 | 8,87 | 7,89 | 6,92 | 5,99 |
| Case 3 | 8,50 | 9,59 | 10,67 | 11,74 | 12,79 | 12,84 | 12,88 | 13,92 | 13,95 | 13,97 | 12,99 | 12,101 | 12,102 | 11,103 | 10,103 | 9,105 | 8,110 |
| Case 4 | 11,66 | 13,78 | 14,88 | 16,96 | 16,103 | 17,109 | 18,114 | 18,118 | 18,121 | 18,123 | 18,125 | 17,126 | 16,126 | 16,125 | 14,124 | 13,124 | 11,125 |
| Case 5 | 15,88 | 18,105 | 20,118 | 22,129 | 23,137 | 24,144 | 25,150 | 25,155 | 25,158 | 25,160 | 25,161 | 24,161 | 23,160 | 22,157 | 20,154 | 18,150 | 15,148 |
| Case 6 | 11,117 | 13,140 | 14,158 | 16,173 | 16,185 | 17,194 | 18,202 | 18,208 | 18,212 | 18,214 | 18,216 | 17,215 | 16,214 | 16,211 | 14,207 | 13,203 | 11,205 |
| Case 7 | 8,158 | 9,190 | 10,215 | 11,235 | 12,252 | 12,265 | 12,275 | 13,283 | 13,289 | 13,293 | 12,295 | 12,294 | 12,293 | 11,290 | 10,288 | 9,288 | 8,301 |
| Case 8 | 5,213 | 6,259 | 7,293 | 8,322 | 8,345 | 9,364 | 9,379 | 9,391 | 9,399 | 9,406 | 9,410 | 9,412 | 8,412 | 8,413 | 7,415 | 6,426 | 5,466 |
| Case 9 | 4,287 | 5,352 | 5,401 | 6,444 | 6,476 | 6,503 | 6,526 | 6,545 | 6,560 | 6,571 | 6,580 | 6,587 | 6,594 | 6,605 | 5,620 | 5,660 | 4,760 |

Table 4
The values of $N^*$, and the minimum cost $TC(N^*)$ for $\lambda = 0.6$

| $\mu$ | 0.7 | 1.05 | 1.4 | 1.75 | 2.1 | 2.45 | 2.8 | 3.15 | 3.5 | 3.85 | 4.2 | 4.55 | 4.9 | 5.25 | 5.6 | 5.95 | 6.3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case 1 | 3,74 | 5,58 | 5,54 | 6,52 | 6,50 | 6,49 | 6,48 | 6,47 | 6,47 | 6,46 | 6,46 | 6,46 | 7,46 | 7,45 | 7,45 | 7,45 | 7,45 |
| Case 2 | 4,8 0 | 6,68 | 7,65 | 8,63 | 8,62 | 9,61 | 9,61 | 9,60 | 9,60 | 9,60 | 9,59 | 9,59 | 9,59 | 9,59 | 9,59 | 9,59 | 9,58 |
| Case 3 | 5,87 | 9,81 | 10,80 | 11,80 | 12,79 | 12,79 | 12,79 | 12,79 | 13,78 | 13,78 | 13,78 | 13,78 | 13,78 | 13,78 | 13,78 | 13,78 | 13,78 |
| Case 4 | 7,98 | 13,100 | 15,102 | 16,103 | 17,104 | 17,104 | 17,104 | 18,104 | 18,105 | 18,105 | 18,105 | 18,105 | 18,105 | 18,105 | 19,105 | 19,105 | 19,105 |
| Case 5 | 10,114 | 18,126 | 21,132 | 22,136 | 23,138 | 24,139 | 25,140 | 25,141 | 25,141 | 25,142 | 26,142 | 26,143 | 26,143 | 26,143 | 26,143 | 26,143 | 26,144 |
| Case 6 | 7,152 | 13,166 | 15,176 | 16,182 | 17,186 | 17,188 | 17,190 | 18,191 | 18,192 | 18,193 | 18,194 | 18,194 | 18,195 | 18,195 | 19,196 | 19,196 | 19,196 |
| Case 7 | 5,216 | 9,225 | 10,239 | 11,247 | 12,253 | 12,256 | 12,259 | 12,261 | 13,263 | 13,264 | 13,265 | 13,266 | 13,267 | 13,268 | 13,268 | 13,269 | 13,269 |
| Case 8 | 4,327 | 6,311 | 7,329 | 8,339 | 8,347 | 9,353 | 9,356 | 9,359 | 9,361 | 9,363 | 9,365 | 9,366 | 9,367 | 9,368 | 9,369 | 9,370 | 9,371 |
| Case 9 | 3,524 | 5,439 | 5,456 | 6,470 | 6,478 | 6,485 | 6,490 | 6,494 | 6,497 | 6,500 | 6,503 | 6,505 | 7,506 | 7,507 | 7,508 | 7,509 | 7,510 |

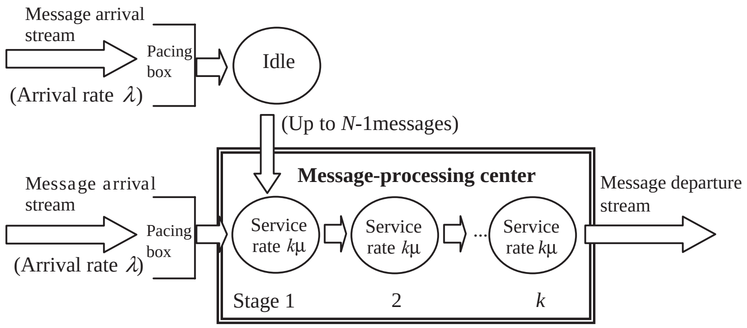

Fig. 3. A flowchart for the message processing center in computer communication networks.

if necessary, receive service from the server, and depart. There is a single “message processing center” (service station) and a “pacing box” (storage room), and the queue size (capacity) is assumed to be infinite. Together, the message processing center and the pacing box mimic the flow control policy, in the following way. The message-processing center does not operate when no messages are present, but may perform an auxiliary task such as maintenance in the so-called idle period with fixed cost. When a source starts sending messages to a particular destination, a pacing count at the source is initialized to the value of zero. This pacing count is incremented every time a message is received. The pacing box stores up to a total of $N$-1 messages. When the $N$th message arrives, it triggers the discharge of the waiting messages into the message-processing center to service. We assume that messages arrive at the pacing box form a single line and wait for service in the order in which they arrive; that is, the first-come first-served discipline. Only a single message could be process at a time. A message request submitted to the system may face two conditions. One is that the message-processing center is busy upon arrival, the message must wait in the pacing box until the processor is available. The other is that the server is in idle status, the message enters the pacing box waiting for counts achieving $N$ to start service.

The message-processing center has $k$ nodes (stages), representing the source node (say stage 1), the destination node (say stage $k$), and $k$-2 intermediate nodes. A single coaxial cable is used to interconnect stations. The processing times of the messages are made up of $k$ independent and identical distributed exponential random variables with mean $1/k\mu$ which yield an Erlang type $k$ distribution. We must specify the workload intensity, which in this case is the rate at which messages arrive (e.g., one message every 2 s or 0.5 messages/s). Arrival process follows a Poisson distribution at a rate $\lambda$. They flow node-to-node, requiring service at each node with mean service rate $k\mu$. As long as its queue is non-empty, it will process message traffic at this rate. The so-called busy period is initiated when the processor starts serving the requests waiting in the system, and terminates when all the requests are served. The next message cannot start processing until the previous message has completed all the stages. The message continues to transmit to the processing center in rate $\lambda$, regardless of the number of outstanding messages; that is, the processing process is independent of the Poisson arrival process of the messages. Our objective is to model the “pacing level“ of messages between a single source/destination pair—the optimum window size $N$ to minimize the total expected cost. A flowchart for the message processing is depicted in Fig. 3.

Table 5
Model input parameter values

| System parameters and cost elements | Notation | Value |
|---|---|---|
| Message stream arrival rate | $\lambda$ | 0.4 |
| Message processing (service) rate | $\mu$ | 1.0 |
| Number of stages (stage parameter) | $k$ | 6 |
| Holding cost per second for each message present in the system | $C_h$ | 5 |
| Cost per second for keeping the server operating | $C_o$ | 50 |
| Start-up cost for turning the server on | $C_s$ | 100 |
| Removable cost per second for removing the service station | $C_r$ | 100 |
| Cost per second for performing an auxiliary task | $C_a$ | 10 |

System characteristics calculations for the model do not require complicated intermediate functions to be implemented, and most of the system performance measures usually of interest can be calculated in a straightforward way. In the example investigated, input system parameters the message stream arrival rate $\lambda = 0.4$ message/s, the message processing (service) rate $\mu = 1.0$ message/s, the number of stages in the Erlang distribution of service-time $k = 6$ and cost element the holding cost per second for each message present in the system set to $C_h = \$5$, the cost per second for keeping the "message-processing center" (service station) operating set to $C_o = \$50$, the start-up cost for turning the "message-processing center" on set to $C_s = \$100$, the removable cost per second for removing the "message-processing center" set to $C_r = \$100$ and cost per second for performing an auxiliary task by the service station set to $C_a = \$10$. The upper bound of $N$ considered is set to $L = 30$ messages. The summary of the model inputs are tabulated in Table 5.

The program output is shown in the following:

$$L_N = 2.0556,$$

$$E[B] = 6.6667,$$

$$E[I] = 10,$$

$$E[C] = 16.6667,$$

$$N = 4,$$

$$TC(N) = 48.2778.$$

The MATLAB computer program gives the expected number of messages in the system $L_N = 2.06$ messages, the expected length of processing (busy) period $E[B] = 6.67$ s, the expected length of idle period $E[I] = 10$ s and the expected length of busy cycle $E[C] = 16.67$ s. The value of $N$ for the optimal management policy, is $N^* = 4$ units, and the corresponding minimum expected cost is found to be $TC(N^*) = \$48.28$. Fig. 4 plots the minimum expected cost $TC(N)$ versus $N = 1(1)30$. The plot shows that the minimum expected cost indeed occurs when $N = 4$, and the tendency of $TC(N)$ versus $N$ could be easily observed. We summarize the model outputs in Table 6. We have given an example to illustrate how a system analyst can use the computer program such as MATLAB to calculate system performance measures, the optimum value of $N$, and its minimum expected cost.

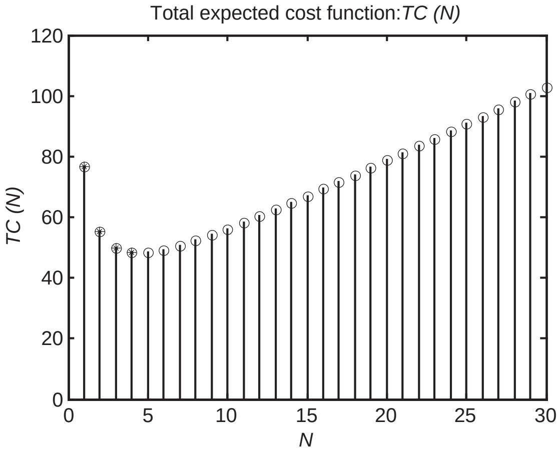

Fig. 4. Plot of $TC(N)$ versus $N$ for $N = 1(1)30$.

Table 6
Model output for system performance measures

| System performance measures | Notation | Value |
|---|---|---|
| Expected number of messages in the system | $L_N$ | 2.06 |
| Expected length of busy period | $E[B]$ | 6.67 |
| Expected length of idle period | $E[I]$ | 10 |
| Expected length of busycycle | $E[C]$ | 16.67 |
| Optimal management policy | $N^*$ | 4 |
| Minimum expected cost | $TC(N^*)$ | 48.28 |

The application example demonstrates the levels of detail that are appropriate for building a model and using that model for performance projection. The example illustrates the relationship between modeling concepts, evaluation algorithms, and modeling software. The example also indicates how such software can save the cost by the analyst.

## References

[1] Yadin M, Naor P. Queuing systems with a removable service station. Operations Research Q 1963;14:393–405.

[2] Bell CE. Characterization and computation of optimal policies for operating an $M/G/1$ queuing system with removable server. Operations Research 1971;19:208–18.

[3] Bell CE. Optimal operation of an $M/G/1$ priority queue with removable server. Operations Research 1972;21: 1281–9.

[4] Heyman DP. Optimal operating policies for $M/G/1$ queuing system. Operations Research 1968;16:362–82.

[5] Kimura T. Optimal control of an $M/G/1$ queuing system with removable server via diffusion approximation. European Journal of Operational Research 1981;8:390–8.

[6] Sobel MJ. Optimal average-cost policy for a queue with start-up and shut-down costs. Operations Research 1969;17:145–58.

[7] Teghem Jr. J. Optimal control of a removable server in an $M/G/1$ queue with finite capacity. European Journal of Operational Research 1987;31:358–67.

[8] Sivazlian BD, Stanfel LE. Analysis of system in operations research. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[9] Gross D, Harris CM. Fundamental of queuing theory, 2nd ed. New York: Wiley, 1985.

[10] Wang KH, Huang HM. Optimal control of an $M/E_k/1$ queuing system with a removable service station. Journal of the Operational Research Society 1995;46:1014–22.

[11] Wang KH, Ke JC. A recursive method to the optimal control of an $M/G/1$ queuing system with finite capacity and infinite capacity. Applied Mathematical Modeling 2000;24:899–914.