Yugoslav Journal of Operations Research
15 (2005), Number 2, 191-207

APPLICATION OF EMPIRICAL BAYESIAN ESTIMATION TO THE OPTIMAL DECISION OF A SERVER-DEPENDENT QUEUING SYSTEM

Pei-Chun LIN

Department of Transportation and Communication Management Science
National Cheng Kung University
Taiwan, R.O.C.
peichunl@gmail.com

Received: November 2003 / Accepted: October 2004

Abstract: This paper presents a decision model that uses empirical Bayesian estimation to construct a server-dependent $M/M/2/L$ queuing system. A Markovian queue with a number of servers depending upon queue length with finite capacity is discussed. This study uses the number of customers for initiating and turning off the second server as decision variables to formulate the expected cost minimization model. In order to conform to the reality, we first collect data of interarrival time and service time by observing a queuing system, then apply the empirical Bayesian method to estimate its traffic intensity. In this research, traffic intensity is used to represent the demand for service facilities. The system initiates another server whenever the number of customers in the system reaches a certain length N and removes the second server as soon as the number of customers in system reduces to Q. Associating the costs with the opening of the second server and the waiting cost of customers, a relationship is developed to obtain the optimal value of N and Q to minimize cost. The mean number of customers in the system and the queue length of customers are derived as the characteristic values of the system. Model development and the implications of the data are discussed in detail.

Keyword: Empirical Bayesian estimation, server-dependent queuing system, traffic intensity.

1. INTRODUCTION

The waiting line of service system is a widespread phenomenon. Customers always wish not to have to wait and to receive service as soon as possible. As customers put a higher value on their time, waiting is regards as a proportionally greater waste.

Hence, managers face the challenge: how to reduce waiting time and achieve customer satisfaction? In order to shorten the wait time, the number of servers must be increased, which at the same time increases the cost of providing services. However, when the demand declines, servers will be idle and resources are wasted, which incur unnecessary cost. It is a critical issue for managers to decide how to allocate servers or resources in an efficient way in order to reduce unnecessary facility cost, idle cost, the cost of losing customers, and to meet the variation of demand.

For instance, the operations of Postal Remittances and Savings Banks (PRSB) in Taiwan face fierce competition under the trends of financial liberalization and internationalization. Customers not only focus on the quality of merchandise but also emphasize the invisible service while making use of postal or financial services. In order to provide better quality of service and reduce customers' waiting time would increase the cost of personnel. Decision makers face the dilemma of obtaining a balance point between providing good quality of service and controlling costs to keep them reasonable. Similarly, the speed of passengers go through an immigration terminal usually influences the reputation of an airport. Travelers' assessment mostly comes from their waiting time. If managers are able to measure the gain and loss between customer waiting and facility costs, it is possible to raise customers' satisfaction and at the same time contain the costs of doing so then they are successful at service facility requirement planning.

The main objective of this study was to establish an evaluation model as a reference for service facility requirement planning. In daily life, it is common to meet all sorts of queues for service, such as the queue for tickets at a cinema, queues of cars waiting to be filled up at a gas station, or even the transfer of network image – all these are situations for the implementation of queuing theory. The number and allocation of servers serving the queue is a problem of service facility requirement planning. In the practical procedure of planning, decision makers may base their plain on the regular flow rate of customers and the expected service rate, or their subjective judgment, to decide the required amount of service requirement and the number of facilities or servers needed. There is a need for an objective and effective model to aid managers to operate systems optimally. This research wishes to implement the empirical Bayesian approach to estimate the service requirement based on the actual operation of queuing. It then constructs a server-dependent queuing system. The controllable system initiates another server whenever the number of customers in system reaches a certain length and turns off the second server whenever the number of customers in system reduces to a certain length. The specific objectives of this research includes:

1. Consider the randomness of customer arrival and service time and incorporate the empirical Bayesian approach to estimate the amount of service required.
2. Construct a server-dependent $M/M/2/L$ queuing system. The system initiates another server whenever the queue length in front of first server reaches a certain length N and closes the second server whenever the queue length in front of first server reduces to a certain length Q. To analyze the system characteristics such as the expected number of customers in system, the probability of server being idle, etc.
3. Use N and Q as decision variables to construct a model to minimize the expected cost associated with the opening of the second server and the waiting of customers.

2. LITERATURE REVIEW

In this section we first explain the reason for using traffic intensity to define the amount of service requirement, then organize how to apply the empirical Bayesian approach to discover the estimator of traffic intensity. Finally we describe the system characteristics and development of a server-dependent queuing system and discuss the related references.

2.1. Traffic intensity vs. the amount of service required

The definition of traffic intensity ρ is the ratio of arrival rate over service rate. It is an important reference of queuing system and represents the utilization or proportion of the server being occupied. This study utilizes traffic intensity as the indication of the amount of service required. The larger traffic intensity means a larger arrival rate or a lower service rate. When $\rho \geq 1$, it means the arrival rate is at least equal to the service rate but it can also exceed the service rate. Obviously a single server is unable to cope with the amount of service requirement. After a period of time, the system will blow up (Winston, 1994). Queues happen due to the uncertainty of the tempo at which customers will be arriving and the variation of service time. There is no waiting time only when customers arrive at a fixed interval and service time is a constant. In reality, customers arrive at random intervals that are unknown in advance and so is the time needed to serve a customer. In order to avoid the assumption that the arrival rate and the service rate as known, this research applies the empirical Bayesian method to estimate the traffic intensity of a queuing system, which can meet the actual randomness and uncertainty and make the model proposed by this study be more reasonable.

2.2. Empirical Bayesian approach

The empirical Bayesian method is based upon a given prior distribution. When a suitable amount of observation values is collected, the prior distribution is used to calculate the posterior distribution. It also applies the concept of maximum likelihood to obtain the estimation of parameters. This section first aims to differentiate the Bayesian and empirical Bayesian methods of estimation, then discusses various methods of statistical analysis for a queuing system, and finally investigates the advantages and adaptability of empirical Bayesian estimation.

Suppose that $X_1,\ldots,X_n$ are independent random variables, each having a probability density function given by

$$g(x_1, x_2 \ldots\ldots x_n \,|\theta) = \prod_{i=1}^{n} f(x_i \,|\theta) \quad \text{if } X_i = x_i,\ i = 1,\ldots,n \ \text{(prior distribution)}$$

where θ is unknown. Further, suppose that θ has the density function $p(\theta)$. The joint distribution of $x_{1,} x_2 \ldots\ldots x_n$ and θ is

$$q(x_1, x_2 \ldots\ldots x_n, \theta) = g(x_1, x_2 \ldots\ldots x_n \,|\theta)\, p(\theta) = \prod_{i=1}^{n} f(x_i \,|\theta)\, p(\theta)$$

The marginal probability density of $x_1, x_2 \ldots\ldots x_n$ is

$$k(x_1, x_2 \ldots\ldots x_n) = \int q(x_1, x_2 \ldots\ldots x_n) d\theta$$

We have the conditional density of θ given $X_1, \ldots, X_n$ is given by

$$h(\theta | x_1, x_2 \ldots\ldots x_n) = \frac{q(x_1, x_2 \ldots\ldots x_n, \theta)}{k(x_1, x_2 \ldots\ldots x_n)} \text{ (posterior density function)}$$

Table 1 illustrates the difference between empirical Bayesian and Bayesian methods in. It shows that the Bayesian method assumes the prior distribution and parameters are known. For the empirical Bayesian method, the compound function of prior distribution is designated such as the most common choice exponential distribution in queuing theory, but the parameters (θ) are unknown.

Table 1: The difference between Bayesian and Empirical Bayesian

Method	Prior distribution	Posterior distribution
Bayesian	$P(\lambda\|\theta)$ Known, θ known	$P(\lambda\|X,\theta) \sim f(X\|\lambda) \cdot P(\lambda\|\theta)$
Empirical Bayesian	$P(\lambda\|\theta)$ Known, θ unknown	$P(\lambda\|X,\hat{\theta}) \sim f(X\|\lambda) \cdot P(\lambda\|\hat{\theta})$

This research used traffic intensity as the indication of the amount of service required. The accuracy of estimation has a major influence on the model of cost analysis constructed subsequently. Mcgrath et al. (1987) applied a Bayesian approach to queuing and pointed out the specification of uncertainty in the estimation of parameters. Thiruvaiyaru et al. (1992) described the advantages of the empirical Bayesian approach on parameter estimation in queuing systems and concluded that the empirical Bayesian approach seeks to combine the logical advantages of the Bayesian techniques with the objective practicality of the frequentist approach.

Other researches that implements empirical Bayesian approach include: Armeto et al. (1994) who emphasized the Bayesian prediction in *M*/*M*/1 queues; Wiper (1998) has implemented empirical Bayesian estimation to Erlang distribution; again Sohn (1996) has concluded that the traffic intensity estimated by empirical Bayesian approach holds the minimal mean square error.

2.3. Server-dependent queues

The major difference of a server-dependent queue from a general queuing system is that the number of servers depends upon the queue length. It was first brought to notice by Singh (1970) that a queuing system could operate in such a way that a new service facility is provided whenever the queue in front of the server reaches a certain

length. Garg et al. (1993) extended the concept and developed the queue $M/M/2$ with a number of homogeneous or heterogeneous servers depending on the queue length. In a two server heterogeneous system, the service rate for the first and the second server are different. They also proposed the conditions for gaining the maximum profit – that the second server should be applied at queue length N. Yamashiro (1996) revised the model of Garg et al. (1993) and assumed that a queue with finite capacity is applicable ($M/M/2/L$). Dai (1999) proposed the finite capacity $M/M/3/L$ queuing system where the number of servers changes depending on the queue length. Bansal et al. (1994) has investigated the factors of cost for activating the second server.

Most of previous researches focused on turning on the second server when queue length reached N. Some of them are set up so that the first server should not be initiated until queue reach length N. Researchers such as Sapna (1996) analyzed the optimal N value for activating the first server under Gamma distribution; Wang et al. (1995) considered the server with unexpected failure to derive the non-reliable $M/M/1/L$ system; Wang et al. (15)[11] drew Erlang distribution into the non-reliable server in a finite and infinite $M/H_2/1$ queuing system. Hsie (1993) took into account that for a $M/M/1$ system, when there is no one to serve, the server would be turned off to reduce idling cost. The above studies all used the optimal queue length N as the decision variable – to decide when to turn on the first server, and constructed the objective function for the minimum expected cost.

Wang et al. (1999) and Dai (1999) added cost in the objective function. This research quantifies customers' waiting cost and considers the cost of activating the second server, and its idle cost to build the model of minimum expected cost. Yamashiro (1996), Wang et al. (1995), Garg et al. (1993) and Dai (1999) didn't describe how to acquire the traffic intensity ρ. This study estimates ρ by the empirical Bayesian approach. Besides, in order to fit the most conditions, we set up a system where the first server is always operating. This study also brings in Wang's (2000) idea and treats the queue length for turning off the second server, as a decision variable.

3. RESEARCH METHOD

This paper applies the empirical Bayesian approach to estimate the demand for service and constructs a server-dependent queuing system, then employs the queue length for activating and closing the second server as decision variables to construct the model of minimum expected cost for a decision maker. In part one we used simulation to produce numerical data or we collect observational data and referred the traffic intensity estimated by the empirical Bayesian approach proposed by Thiruvaiyaru (1992) to indicate of the required amount of service. Next, we derived the probabilities of each state for a M/M/2/L server-dependent queuing system. Finally, we add in the parameters of cost and combine the first two parts to solve the optimal queue length N for starting a second server and the optimal queue length Q for turning off the second server. We first introduced the method to apply the empirical Bayesian approach and obtain observational data to generate the estimation of traffic intensity.

3.1. Empirical Bayesian estimator of traffic intensity

Thiruvaiyaru (1992) supposed there are *H* independent *M*/*M*/1 queues in which the interarrival times $\{U_{ik}, i=1,...,n\}$ of the first n customer, and the service times $\{V_{jk},\ j=1,...,m\}$ of the first m customers are observed for $k=1,...,H$. Given the arrival rate λ_k, $\{U_{ik}, k=1,...,H\}$ are i.i.d exponential (λ_k) random variables; that is

$$f_{U_k}(u_k|\lambda_k)=\lambda_k^n \exp\{-\lambda_k \sum_{i=1}^{n} u_{ik}\}$$

where

$$\mathbf{U}_k=(U_{ik}, i=1,...,n)'$$

Also, given the service rate μ_k, $\{V_{jk}, j=1,...,m\}$ are i.i.d. exponential (μ_k) random variables; that is,

$$f_{\mathbf{V}_k}(v_k|\mu_k)=\mu_k^m \exp\{-\mu_k \sum_{j=1}^{m} v_{jk}\}$$

where

$$\mathbf{V}_k=(V_{jk},\ j=1,...,m)'$$

The arrival rates $\{\lambda_1,...,\lambda_N\}$ are assumed to be i.i.d. $Gamma(\alpha_1,\beta_1)$ (prior distribution) and the service rates $\{\mu_1,...,\mu_k\}$ are assumed to be i.i.d. $Gamma(\alpha_2,\beta_2)$ (prior distribution). Also, the two sequences $\{\lambda_1,...,\lambda_N\}$ and $\{\mu_1,...,\mu_k\}$ are assumed to be independent of each other. The empirical Bayesian estimator is derived as

$$\hat{\rho}^{EB}=\frac{(n+\hat{\alpha}_1)(\sum_{j=1}^{m} V_j+\hat{\beta}_2)}{(m+\hat{\alpha}_2-1)(\sum_{i=1}^{n} U_i+\hat{\beta}_1)}$$

where $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2$ are the one-step maximum likelihood estimators of $\alpha_1, \alpha_2, \beta_1, \beta_2$, respectively. First, let $\hat{\boldsymbol{\eta}}_l=(\hat{\alpha}_l, \hat{\beta}_l)', l=1,2$ be the one-step Maximum likelihood estimators of $\boldsymbol{\eta}_l=(\alpha_l,\beta_l)',\ l=1,2$, respectively. Let $m_{11}=\sum_{k=1}^{H}\sum_{i=1}^{n}\frac{U_{ik}}{Hn}$ and $m_{21}=\sum_{k=1}^{H}\sum_{i=1}^{n}\frac{U_{ik}^2}{Hn}$, we can calculate $m_{11}=\beta_1/(\alpha_1-1)$ and $m_{21}=2\beta_1^2/(\alpha_1-1)(\alpha_1-2)$. The moment estimators $(\tilde{\alpha}_1,\tilde{\beta}_1)$ of α_1, β_1 are

$$\tilde{\alpha}_1=2(m_{21}-m_{11}^2)/m_{21}-2m_{11}^2)$$

$$\tilde{\beta}_1=m_{11}m_{21}/(m_{21}-2m_{11}^2)$$

Again, let $m_{12} = \sum_{k=1}^{H}\sum_{j=1}^{m} V_{jk} \big/ (Hm)$ and $m_{22} = \sum_{k=1}^{H}\sum_{j=1}^{m} V_{jk}^2 \big/ (Hm)$, we obtain the moment estimator of (α_2, β_2):

$$\tilde{\alpha}_2 = 2(m_{22} - m_{12}^2) \big/ (m_{22} - 2m_{12}^2)$$

$$\tilde{\beta}_2 = m_{22} m_{12} \big/ (m_{22} - 2m_{12}^2)$$

Then, the one-step maximum likelihood estimators of $\boldsymbol{\eta}_l = (\alpha_l, \beta_l)'$, $l = 1, 2$ are given by

$$\hat{\boldsymbol{\eta}}_l = \tilde{\boldsymbol{\eta}}_l - W_l^{-1}(\tilde{\boldsymbol{\eta}}) \cdot S_l(\tilde{\boldsymbol{\eta}}), \; l = 1, 2$$

where the marginal likelihood function is

$$L = f(x_1, ..., x_n) = \prod_{k=1}^{H} f(\mathbf{U}_k) f(\mathbf{V}_k)$$

$$= \prod_{k=1}^{H} \left[\frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_1 + n)}{(\sum_{i=1}^{n} \mu_{ik} + \beta_1)^{\alpha_1 + n}} \cdot \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} \cdot \frac{\Gamma(\alpha_2 + m)}{(\sum_{j=1}^{m} v_{jk} + \beta_2)^{m + \alpha_2}} \right]$$

and

$$\tilde{\boldsymbol{\eta}}_l = (\tilde{\alpha}_l, \tilde{\beta}_l)', \; l = 1, 2$$

and

$$S_l(\tilde{\boldsymbol{\eta}}) = \left[\frac{\partial \ln L}{\partial \alpha_1}, \quad \frac{\partial \ln L}{\partial \beta_l} \right]'_{\eta_l = \tilde{\eta}_l}, \; l = 1, 2$$

and

$$W_l(\tilde{\boldsymbol{\eta}}) = \begin{bmatrix} \dfrac{\partial^2 \ln L}{\partial \alpha_l^2} & \dfrac{\partial^2 \ln L}{\partial \alpha_l \partial \beta_l} \\ \dfrac{\partial^2 \ln L}{\partial \alpha_l \partial \beta_l} & \dfrac{\partial^2 \ln L}{\partial \beta_l^2} \end{bmatrix}_{\eta_l = \tilde{\eta}_l}, \; l = 1, 2$$

3.2. Server-dependent M/M/2/L queuing system

The major objective of this section is to establish a server-dependent *M*/*M*/2/*L* queuing system with finite capacity *L*. This system has been set up so that the first server is always on. When the number of customers in the system reaches *N*, the second sever would be activated to release the congestion in the system; when the number of customers in systems reduces to *Q*, it signifies the status of overcrowding has ceased so we can turn off the second server to cut cost. The number of waiting line is only one as shown in Figure 1.

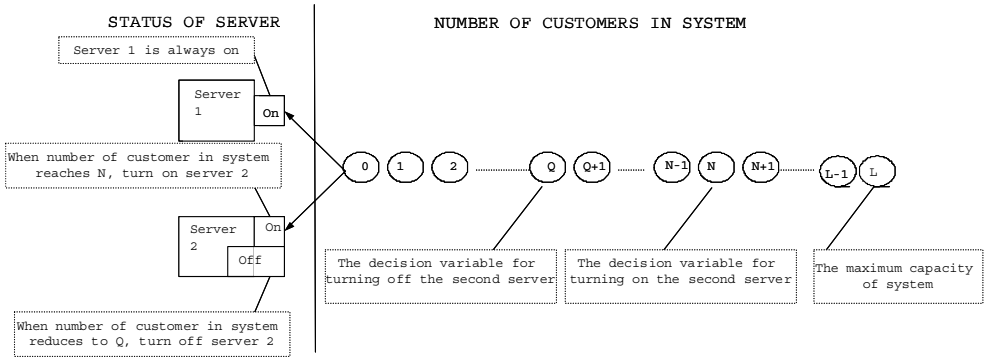

Figure 1: Server-dependent queuing system with single waiting line

The assumptions, parameters and variables used in the model are defined as follows:

Assumptions:

1. The service rule is FCFS.
2. The interarrival time of customers is assumed to be exponential distribution with unknown parameters.
3. The service time for each customer is assumed to be exponential distribution with unknown parameters.
4. The service system could provide two servers at most, but at least one server should remain on to serve customers.
5. The system has finite capacity L and $L>>N$.
6. The service rates of two servers are identical.
7. $1<\rho<2$.

Definition of symbols

1. λ: arrival rate of customers
2. μ: service rate of server
3. ρ: traffic intensity $=\dfrac{\lambda}{\mu}$
4. i: number of servers in service, $i=1,2$
5. j: number of customers in system, $j=0...L$
6. $P(1,j)$: the steady-state probability of only one server is providing service as the number of customers in system is j, where $j=0,1,2,...,Q,Q+1,...,N-1$
7. $P(2,j)$: the steady-state probability of two servers are both providing service as the number of customers in system is j, where $j=Q+1,Q+2,...,N,N+1,...L-1,L$

Based upon the above assumptions and symbols, this research constructed a server-dependent *M*/*M*/2/*L* system. The rate diagram of birth and death process is shown as Figure 2 and the flow balance equations are as follows:

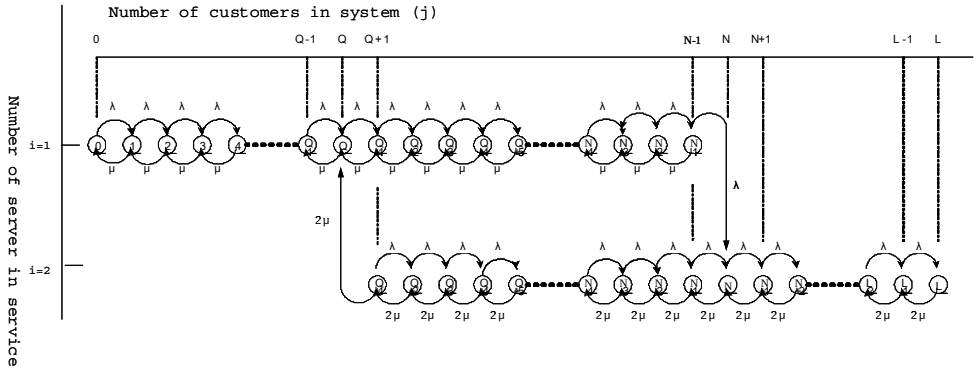

Figure 2: Rate diagram for M/M/2/L queuing system

$$\lambda P(1,0) = \mu P(1,1)$$

$$(\lambda+\mu)P(1,j) = \lambda P(1,j-1) + \mu P(1,j+1) \qquad \text{where } 1 \le j \le Q-1$$

$$(\lambda+\mu)P(1,Q) = \lambda P(1,Q-1) + \mu P(1,Q+1) + 2\mu P(2,Q+1)$$

$$(\lambda+\mu)P(1,j) = \lambda P(1,j-1) + \mu P(1,j+1) \qquad \text{where } Q+1 \le j \le N-2$$

$$(\lambda+\mu)P(1,N-1) = \lambda P(1,N-2)$$

$$(\lambda+2\mu)P(2,Q+1) = 2\mu P(2,Q+2)$$

$$(\lambda+2\mu)P(2,j) = \lambda P(2,j-1) + 2\mu P(2,j+1) \qquad \text{where } Q+2 \le j \le N-1$$

$$(\lambda+2\mu)P(2,N) = \lambda P(2,N-1) + 2\mu P(2,N+1) + \lambda P(1,N-1)$$

$$(\lambda+2\mu)P(2,j) = \lambda P(2,j-1) + 2\mu P(2,j+1) \qquad \text{where } N+1 \le j \le L-1$$

$$\lambda P(2,L-1) = 2\mu P(2,L)$$

To solve the above birth-death flow balance equations, we begin by expressing all the $P(1,j)$'s and $P(2,j)$'s in terms of $P(1,0)$.

1. $i=1$ (only one server is providing service)

$$P(1,j) = P(1,0), \quad j=0$$

$$P(1,j) = \rho^j \cdot P(1,0)_{\text{¡}}, 1 \le j \le Q \tag{1}$$

$$P(1,j) = \frac{\rho \cdot (\rho^{j-1} - \rho^{N-1})}{(1-\rho^{N-Q})} \cdot P(1,0)_{\text{¡}}, \quad Q+1 \le j \le N-1 \tag{2}$$

2. $i = 2$ (two servers both provide service)

$$P(2, j) = \frac{\rho^N \cdot (1-\rho) \cdot [1-(\frac{\rho}{2})^{j-Q}]}{(2-\rho) \cdot (1-\rho^{N-Q})} \cdot P(1,0)¡, \ Q+1 \le j \le N \tag{3}$$

$$P(2, j) = \frac{\rho^N \cdot (1-\rho) \cdot [1-(\frac{\rho}{2})^{N-Q}] \cdot (\frac{\rho}{2})^{j-N}}{(2-\rho) \cdot (1-\rho^{N-Q})} \cdot P(1,0)¡, \ N+1 \le j \le L \tag{4}$$

3. The steady-state probabilities must sum to 1

$$\sum_{i=1}^{2} \sum_{j=0}^{L} P(i, j) = 1 \tag{5}$$

Substituting (1), (2), (3), and (4) into (5) yields

$$\sum_{j=0}^{Q} \rho^j \cdot P(1,0) + \sum_{j=Q+1}^{N-1} \frac{\rho \cdot (\rho^{j-1} - \rho^{N-1})}{(1-\rho^{N-Q})} \cdot P(1,0)$$

$$+ \sum_{j=Q+1}^{N} \frac{\rho^N \cdot (1-\rho) \cdot [1-(\frac{\rho}{2})^{j-Q}]}{(2-\rho) \cdot (1-\rho^{N-Q})} \cdot P(1,0)$$

$$+ \sum_{j=N+1}^{L} \frac{\rho^N \cdot (1-\rho) \cdot [1-(\frac{\rho}{2})^{N-Q}] \cdot (\frac{\rho}{2})^{j-N}}{(2-\rho) \cdot (1-\rho^{N-Q})} \cdot P(1,0) = 1$$

Thus

$$P(1,0) \cdot \left\{ \frac{1}{1-\rho} - \frac{\rho^N \cdot \{(2-\rho) \cdot (N-Q) + \rho \cdot (1-\rho) \cdot (\frac{\rho}{2})^{L-N} \cdot [1-(\frac{\rho}{2})^{N-Q}]\}}{(2-\rho)^2 \cdot (1-\rho^{N-Q})} \right\} = 1$$

We can solve for $P(1,0)$, which is the steady-state probability of no customer in the system:

$$P(1,0) = \left\{ \frac{1}{1-\rho} - \frac{\rho^N \cdot \{(2-\rho) \cdot (N-Q) + \rho \cdot (1-\rho) \cdot (\frac{\rho}{2})^{L-N} \cdot [1-(\frac{\rho}{2})^{N-Q}]\}}{(2-\rho)^2 \cdot (1-\rho^{N-Q})} \right\}^{-1} \tag{6}$$

Then (6) can be used to determine $P(1, j)$, $P(2, j)$. Each of them is a function of traffic intensity ρ, and the decision variables *N*, *Q*. Now we can incorporate the parameters of costs and formulate an NLP to minimize the sum of expected costs due to customer waiting and server operating.

Formulation of objective function

Next we construct an objective function of minimizing expected cost for the *M*/*M*/2/*L* controllable queuing system. The definitions of parameters are as follows:

Ec : expected cost

C_s : the fulltime operating cost for second server

C_i : the fulltime idle cost for second server

C_L : the penalty cost for system being fully loaded

C_e : the penalty cost for system being empty

C_{on} : the start up cost for turning the second server on back and forth

C_{off} : the shut down cost for turning the second server off back and forth

C_w : the average waiting cost for each customer (we assume the expected waiting cost is proportional to the queue length)

The expected cost function is given by

$$\begin{aligned} Ec(N,Q|\rho) = {} & C_s \cdot \sum_{Q+1}^{L} P(2,j) \\ & + C_w \cdot \sum_{j=0}^{N-1} Max[0,(j-1)] \cdot P(1,j) \\ & + C_w \cdot \sum_{j=Q+1}^{L} Max[0,(j-2)] \cdot P(2,j) \\ & + C_i \cdot \sum_{j=0}^{N-1} P(1,j) \\ & + C_{on} \cdot P(2,N) \\ & + C_{off} \cdot P(1,Q) \\ & + C_e \cdot P(1,0) \\ & + C_L \cdot P(2,L) \end{aligned} \tag{7}$$

Next substituting $i=1$ into (1) and (2) yields the sum of probability of one server ($P(1,j)$, $j=0,\ldots,Q,\ldots,N-1$) in system:

$$\sum_{j=0}^{N-1} P(1,j) = \sum_{j=0}^{Q} P(1,j) + \sum_{j=Q+1}^{N-1} P(1,j) = [\sum_{j=0}^{Q} \rho^j \cdot + \sum_{j=Q+1}^{N-1} \frac{\rho \cdot (\rho^{j-1} - \rho^{N-1})}{(1-\rho^{N-Q})}] \cdot P(1,0)$$

where

$$\sum_{j=0}^{Q} \rho^j \cdot P(1,0) = \frac{\rho^{Q+1}-1}{\rho-1} \cdot P(1,0)$$

$$\sum_{j=Q+1}^{N-1} \frac{\rho \cdot (\rho^{j-1} - \rho^{N-1})}{(1-\rho^{N-Q})} \cdot P(1,0) = \frac{[(N-Q-1) \cdot \rho^{Q+N+1} + (Q-N) \cdot \rho^{Q+N} + \rho^{2Q+1}]}{(\rho-1)(\rho^N - \rho^Q)} \cdot P(1,0)$$

Then substituting $i=1$ into (1) and (2) yields the sum of $j \cdot P(1, j)$, $j=0,\ldots,Q,\ldots,N-1$:

$$\sum_{j=0}^{N-1} j \cdot P(1, j) = \sum_{j=0}^{Q} j \cdot P(1, j) + \sum_{j=Q+1}^{N-1} j \cdot P(1, j)$$

$$= [\sum_{j=0}^{Q} j \cdot \rho^{j} \cdot + \sum_{j=Q+1}^{N-1} j \cdot \frac{\rho \cdot (\rho^{j-1} - \rho^{N-1})}{(1-\rho^{N-Q})}] \cdot P(1,0)$$

where

$$\sum_{j=0}^{Q} j \cdot \rho^{j} \cdot P(1,0) = \frac{[Q \cdot \rho^{Q+2} - (Q+1) \cdot \rho^{Q+1} + \rho]}{(\rho-1)^{2}} \cdot P(1,0)$$

and

$$\sum_{j=Q+1}^{N-1} \frac{\rho \cdot (\rho^{j-1} - \rho^{N-1})}{(1-\rho^{N-Q})} \cdot P(1,0) =$$

$$= \frac{\rho^{(Q+N+1)} N - \rho^{(Q+N)} N + \rho^{(2Q+1)} - \rho^{(Q+N+1)} Q - \rho^{(Q+N+1)} + \rho^{(Q+N)} Q}{(-\rho^{Q} + \rho^{N})(\rho-1)} \cdot P(1,0)$$

Substituting $i=2$ into (3) and (4) yields the sum of probability $P(2, j)$, for $j=Q+1,\ldots,N,N+1,\ldots,L$

$$\sum_{j=Q+1}^{L} P(2, j) = \sum_{j=Q+1}^{N} P(2, j) + \sum_{j=N+1}^{L} P(2, j)$$

where

$$\sum_{j=Q+1}^{N} P(2, j) = \sum_{j=Q+1}^{N} \frac{\rho^{N} \cdot (1-\rho) \cdot [1-(\frac{\rho}{2})^{j-Q}]}{(2-\rho) \cdot (1-\rho^{N-Q})} \cdot P(1,0)¡$$

$$= \frac{(-\rho N - \rho + 2N + 2^{(-N+Q)} \rho^{(N+1-Q)} + \rho Q - 2Q) \rho^{N} (1-\rho)}{(\rho-2)(-\rho+2)(1-\rho^{(N-Q)})} \cdot P(1,0)$$

and

$$\sum_{j=N+1}^{L} P(2, j) =$$

$$= \sum_{j=N+1}^{L} \frac{\rho^{N} \cdot (1-\rho) \cdot [1-(\frac{\rho}{2})^{N-Q}] \cdot (\frac{\rho}{2})^{j-N}}{(2-\rho) \cdot (1-\rho^{N-Q})} \cdot P(1,0)¡$$

$$= \frac{(\rho-1)(\rho^{(Q+L+1)} 2^{(N-L)} - \rho^{(Q+N+1)} - 2^{(Q-L)} \rho^{(N+L+1)} + 2^{(-N+Q)} \rho^{(2N+1)}}{(\rho-2)^{2} (-\rho^{Q} + \rho^{N})} \cdot P(1,0)$$

Then substituting $i=2$ into (3) and (4) yields the sum of $j\cdot P(2,j)$, for $j=Q+1,\ldots,N,N+1,\ldots,L$

$$\sum_{j=Q+1}^{L} j\cdot P(2,j)=\sum_{j=Q+1}^{N} j\cdot P(2,j)+\sum_{j=N+1}^{L} j\cdot P(2,j)$$

where

$$\sum_{j=Q+1}^{N} j\cdot P(2,j)=j\cdot\sum_{j=Q+1}^{N}\frac{\rho^{N}\cdot(1-\rho)\cdot[1-(\frac{\rho}{2})^{j-Q}]}{(2-\rho)\cdot(1-\rho^{N-Q})}\cdot P(1,0)$$

$$=\frac{\left[\begin{array}{l}(8N+8N^2-8Q^2+4)\rho^{N+Q+1}+(2^{Q-N+2}+6N\cdot 2^{Q-N})\rho^{2N+2}\\+4(Q^2+Q-N^2-N)\rho^{N+Q}+(5Q^2-Q-5N-5N^2-4)\rho^{N+Q+2}\\+(1-N)\cdot 2^{Q-N+2}\cdot\rho^{2N+1}+(Q-Q^2+N+N^2)\rho^{N+Q+3}\\-2^{Q+1-N}\cdot N\cdot\rho^{2N+3}\end{array}\right]}{2\cdot(\rho-2)^3\cdot(\rho^{Q}-\rho^{N})}\cdot P(1,0)$$

and

$$\sum_{j=N+1}^{L} j\cdot P(2,j)=j\frac{\rho^{N}\cdot(1-\rho)\cdot[1-(\frac{\rho}{2})^{N-Q}]\cdot(\frac{\rho}{2})^{j-N}}{(2-\rho)\cdot(1-\rho^{N-Q})}\cdot P(1,0)$$

$$=(\rho-1)\cdot\frac{\left[\begin{array}{l}(L+1)\cdot 2^{Q-L+1}\rho^{Q-L+1}-(L+1)\cdot 2^{N-L+1}\rho^{Q+L+1}+2^{N-L}\cdot L\cdot\rho^{2+Q+L}\\-2^{Q-L}\cdot L\cdot\rho^{2+L+N}+2(1+N)\rho^{Q+N+1}-2^{Q-N+1}(1+N)\rho^{2N+1}\\-N\rho^{N+Q+2}+2^{Q-N}N\cdot\rho^{2+2N}\end{array}\right]}{(\rho-2)^3\cdot(\rho^{Q}-\rho^{N})}\cdot P(1,0)$$

Then we rewrite (7) as function of traffic intensity ρ, and decision variables *N*, *Q*. ρ is estimated by empirical Bayesian estimator $\hat{\rho}^{EB}$ and substituting $\hat{\rho}^{EB}$ into (6), we obtain $\hat{P}(1,0)$:

$$\hat{P}(1,0)=\left\{\frac{1}{1-\hat{\rho}^{EB}}-\frac{\hat{\rho}^{EBN}\cdot\{(2-\hat{\rho}^{EB})\cdot(N-Q)+\hat{\rho}^{EB}\cdot(1-\hat{\rho}^{EB})\cdot(\frac{\hat{\rho}^{EB}}{2})^{L-N}\cdot[1-(\frac{\hat{\rho}^{EB}}{2})^{N-Q}]\}}{(2-\hat{\rho}^{EB})^2\cdot(1-(\hat{\rho}^{EB})^{N-Q})}\right\}^{-1}$$

The expected cost minimization model is as follows:

$$
\begin{aligned}
Minimize(N,Q\,|\,\hat{\rho}^{EB}) = {} & C_s \cdot \sum_{j=Q+1}^{N} \frac{(\hat{\rho}^{EB})^N \cdot (1-\hat{\rho}^{EB}) \cdot [1-(\frac{\hat{\rho}^{EB}}{2})^{j-Q}]}{(2-\hat{\rho}^{EB}) \cdot [1-(\hat{\rho}^{EB})^{N-Q}]} \cdot \hat{P}(1,0) \\
& + C_s \cdot \sum_{j=N+1}^{L} \frac{(\hat{\rho}^{EB})^N \cdot (1-\hat{\rho}^{EB}) \cdot [1-(\frac{\hat{\rho}^{EB}}{2})^{N-Q}] \cdot (\frac{\hat{\rho}^{EB}}{2})^{j-N}}{(2-\hat{\rho}^{EB}) \cdot [1-(\hat{\rho}^{EB})^{N-Q}]} \cdot \hat{P}(1,0) \\
& + C_w \cdot \sum_{j=0}^{Q} Max[0,(j-1)] \cdot (\hat{\rho}^{EB})^j \cdot \hat{P}(1,0) \\
& + C_w \cdot \sum_{j=Q+1}^{N-1} Max[0,(j-1)] \cdot \frac{\hat{\rho}^{EB} \cdot [(\hat{\rho}^{EB})^{j-1} - (\hat{\rho}^{EB})^{N-1}]}{[1-(\hat{\rho}^{EB})]^{N-Q}} \cdot \hat{P}(1,0) \\
& + C_w \cdot \sum_{j=Q+1}^{N} Max[0,(j-2)] \cdot \frac{(\hat{\rho}^{EB})^N \cdot (1-\hat{\rho}^{EB}) \cdot [1-(\hat{\rho}^{EB}/2)^{j-Q}]}{(2-\hat{\rho}^{EB}) \cdot (1-(\hat{\rho}^{EB})^{N-Q})} \cdot \hat{P}(1,0) \\
& + C_w \cdot \sum_{j=N+1}^{L} Max[0,(j-2)] \cdot \frac{(\hat{\rho}^{EB})^N \cdot (1-\hat{\rho}^{EB}) \cdot [1-(\hat{\rho}^{EB}/2)^{N-Q}]}{(2-\hat{\rho}^{EB}) \cdot (1-(\hat{\rho}^{EB})^{N-Q})} \cdot (\frac{\hat{\rho}^{EB}}{2})^{j-N} \cdot \hat{P}(1,0) \\
& + C_i \cdot \left[\sum_{j=1}^{Q} (\hat{\rho}^{EB})^j \cdot \hat{P}(1,0) + \sum_{j=Q+1}^{N-1} \frac{\hat{\rho}^{EB} \cdot [(\hat{\rho}^{EB})^{j-1} - (\hat{\rho}^{EB})^{N-1}]}{[1-(\hat{\rho}^{EB})]^{N-Q}} \cdot \hat{P}(1,0) \right] \\
& + C_{on} \cdot \frac{(\hat{\rho}^{EB})^N \cdot (1-\hat{\rho}^{EB}) \cdot [1-(\hat{\rho}^{EB}/2)^{N-Q}]}{(2-\hat{\rho}^{EB}) \cdot (1-(\hat{\rho}^{EB})^{N-Q})} \cdot \hat{P}(1,0) \\
& + C_{off} \cdot (\hat{\rho}^{EB})^Q \cdot \hat{P}(1,0) \\
& + C_e \cdot \hat{P}(1,0) + C_L \cdot \frac{(\hat{\rho}^{EB})^N \cdot (1-\hat{\rho}^{EB}) \cdot [1-(\hat{\rho}^{EB}/2)^{N-Q}]}{(2-\hat{\rho}^{EB}) \cdot (1-(\hat{\rho}^{EB})^{N-Q})} \cdot (\frac{\hat{\rho}^{EB}}{2})^{L-N} \cdot \hat{P}(1,0)
\end{aligned}
$$

It is hard to solve the above NLP analytically and prove its feasible region is a convex set which possesses the optimal N^* and Q^* and minimizes the expected cost globally. Thus, we applied a numerical method to explore how changes in the NLP's parameters change the optimal solution.

4. SENSITIVITY ANALYSIS

In this section we illustrate some the results obtained in previous sections with a hypothetical queuing experiment. First we applied Monte Carlo simulation to generate random data for five queues and the one-step maximum likelihood estimator $(\hat{\alpha}_1, \hat{\beta}_1)$ = (58.19542203, 11.31472722), $(\hat{\alpha}_2, \hat{\beta}_2)$ = (28.31890446, 6.954543058). The empirical Bayesian estimator of traffic intensity $\hat{\rho}^{EB}$ =1.294695872.

Next, we perform numerical analysis to determine:

- The influence of changing C_s on the minimum expected cost and optimal N^*, Q^* as $C_{on} = C_{off}$ = 0 and $C_{on} = C_{off}$ = 25, respectively.

- The impact of changing C_i on the minimum expected cost and optimal N^*, Q^* as $C_{on} = C_{off} = 0$ and $C_{on} = C_{off} = 25$, respectively.
- The impact of changing the average waiting cost for each customer C_w on the minimum expected cost and optimal N^*, Q^* as $C_{on} = C_{off} = 25$
- The influence of C_L and C_e on the minimum cost respectively.
- The impact of changing C_{on} and C_{off} on the minimum expected cost and optimal N^*, Q^* while $C_s = 0$, $C_e = 0$ and $C_s = 150$, $C_e = 250$, respectively.
- The influence of traffic intensity on the optimal solution.
- The influence of system capacity L on the optimal solution.

We sum up the following results:

- When there is no start up and shut down cost for the second server, in order to attain the minimum cost the second server will be turned on and off frequently. As the fulltime operating cost for second server gets higher, the second server won't provide service readily.
- When the fulltime idle cost for second server gets larger, the second server should be kept busy most of the time. As long as the average waiting cost for each customer becomes larger, the system had better not to keep customer wait so the second server should be turns on sooner.
- We found the variation of penalty cost for system being fully loaded and empty reveal no significant impact on the minimum cost. However, the penalty cost for system being fully empty did change optimal N^* and Q^* significantly.
- If there were no cost for the second server to offer service, the second server would be turned on as soon as possible. However, the start up cost and shut down cost would prevent the second server from being turned on and off.
- The higher the traffic intensity is, the sooner the second server should be turned on to cease the congestion.
- When the system capacity L is big enough, it makes no influence on the optimal solution.

Table 2 presents the special case in which only one parameter is non-zero to validate the accuracy of the proposed model. Finally we sum up the effect of increasing parameters on the decision variables in Table 3.

Table 2: The special case in which only one parameter is non-zero

$\hat{\rho}^{EB}$ =1.2947, L=50									
C_s	C_i	C_w	C_L	C_e	C_{on}	C_{off}	N^*	Q^*	Ec
150	0	0	0	0	0	0	49	47	32.22
0	100	0	0	0	0	0	2	0	43.02
0	0	1	0	0	0	0	2	0	0.76
0	0	0	500	0	0	0	2	0	≈ 0.00
0	0	0	0	500	0	0	49	47	≈ 0.00
0	0	0	0	0	100	100	49	0	2.16

Table 3: The effect of increasing parameters on the decision variables

Parameter	C_s		C_i		C_w		C_{on} , C_{off}		C_e		C_L	
Decision variable	N^*	Q^*	N^*	Q^*	N^*	Q^*	N^*	Q^*	N^*	Q^*	N^*	Q^*
Effect	+	+	-	-	-	-	+	-	+	+	-	-

5. CONCLUSIONS

This paper applies the empirical Bayesian approach to estimate the demand for service and constructs a server-dependent queuing system, then employs the queue length as decision variables for activating and closing the second server to construct an NLP model of minimum expected cost for a decision maker. Associating the costs with the opening of the second server, the start up and shut down cost for turning on and off the second server, and the waiting of the customers, a relationship is developed to obtain the optimal value of *N* and *Q* to minimize cost. We also performed a sensitivity analysis to discuss how changes in the NLP's parameters (C_s ; C_i ; C_L ; C_e ; C_{on} ; C_{off} ; C_w) affect the optimal solution. From the numerical analysis, we conclude that (1) the fulltime operating cost for the second server and the penalty cost for system being empty increase would cause larger *N** and *Q**; (2) the fulltime idle cost for the second server, the average waiting cost for each customer, and the penalty cost for system being fully loaded increase would cause smaller *N** and *Q**; (3) the start up and shut down cost for turning the second server on and off back and forth increase would cause larger *N** but smaller *Q**. The results of the evaluation model present a reference for service facility requirement planning.

REFERENCES

[1] Armero, C., and Bayarri, M.J., "Bayesian prediction in *M*/*M*/1 queues", *Queueing Systems*, 15 (1994) 401-417.

[2] Bansal, K.K., and Garg, R.L., "An additional space special service facility heterogeneous queue", *Microelectronics and Reliability*, 35(4) (1994) 725-730.

[3] Dai, K.Y., "Queue-dependent servers in an *M*/*M*/3 queueing system with finite capacity", Master Thesis, National Chung Hsing University, Taiwan, 1999.

[4] Garg, R.L., and Singh, P., "Queue dependent servers queuing system", *Microelectronics and Reliability*", 33(15) (1993) 2289-2295.

[5] Hsieh, W.F., "Optimal control of the finite capacity and infinite capacity with a removable service station subject to breakdown", Master Thesis, National Chung Hsing University, Taiwan, 1993.

[6] Mcgrath, M.F., and Gross, D., "A subjective Bayesian approach to the theory of queues I-modeling", *Queueing Systems*, 1 (1987) 317-333.

[7] Sapna, K.P., "An M/G/1-type queueing system with non-perfect servers and no waiting capacity", *Microelectronics and Reliability*, 36(5) (1996) 697-700.

[8] Singh, V.P., "Two-server Markovian queues with balking. Heterogeneous vs. homogeneous servers", *Operations Research*, 18(1) (1970) 145-59.

[9] Sohn, S.Y., ”Influence of a prior distribution on traffic intensity estimation with covariate”, *Journal of Statistical Computation & Simulation*, 55 (1996) 169-180.

[10] Thiruvaiyaru, D., and Basawa, I.V., “Empirical Bayes estimation for queueing systems and networks”, *Queueing Systems*, 11 (1992) 179-202.

[11] Wang, K.H., and Huang, H.M., “Optimal control of a removable server in an $M/E_k/1$ queueing system with finite capacity”, *Microelectronics and Reliability*, 35(7) (1995) 1023-1030.

[12] Wang, K.H., and Hsieh, W.F., “Optimal control of a removable and non-reliable server in a Markovian queueing systems with finite capacity”, *Microelectron. Reliab.*. 35(2) (1995) 189-196.

[13] Wang, K.-H., Chang, K.-W., and Sivazlian, B.D., “Optimal control of a removable and non-reliable server in an infinite and a finite $M/H_2/1$ queueing system”, *Applied Mathematical Modelling*, 23(8) (1999) 651-666.

[14] Wang, Y.L., "Optimal control of an M/M/2 queueing system with finite capacity operating under the triadic (0,Q,N,M) policy", Master Thesis, National Chung Hsing University. Taiwan, 2001.

[15] Winston, W. L., *Operations Research*, 3rd edition, Duxbury, Indiana University, 1994.

[16] Wiper, M.P., “Bayesian analysis of $E_r/M/1$ and $E_r/M/c$ Queues”, *Journal of Statistical Planning and Influence*, 69 (1998) 65-79.

[17] Yamashiro, M., “A system where the number of servers changes depending on the queue length”, *Microelectronics and Reliability*, 36(3) (1996) 389-391.

[18] Yen, K.L., "Optimal control of the M/Hk/1 queueing system with a single removable server", Master Thesis, National Chung Hsing University, Taiwan, 2000.