AN ALTERNATIVE MEASURE FOR EVALUATING THE DELAY

Hsing Luh*
Department of Mathematical Sciences
National Chengchi University
64, Section 2, Chih-nan Road, Wenshan, Taipei, Taiwan, 116, R.O.C.

ABSTRACT

Queues that involve waiting as a model for analyzing the possible delay have been studied for past several decades. The average total time spent in the system and the average total time spent in queue are two of the most fundamental quantities describing a queueing system's behavior. In addition, there are other important factors that will affect the delay as well. For example, the environment in which queue waiting occurs plays a fundamental role in a customer's perceived and/or actual cost of participating in that system. In this paper, we study a queueing model that takes the waiting environment into account. We present a very simple solution procedure to solve the steady-state probability in this model with the required measurement.

Keywords: delay, queues, M/G/1.

1. INTRODUCTION

Studying a queueing model or a queueing network model that take a possible delay into account in a system has been a considerably long history because both models have significant representation as been used for the design and analysis of many applied systems. Examples may be found in Robertazzi [12], Buzacott and Shanthikumar [2]. Traditionally, the Little's formula, $L = \lambda W$, represents the relationship between the time-average number of customers in the system L and the average total time spent in the system W while λ is the arrival rate per unit time. In general, the average delay, namely the average total time either spent in the system or in the queue, has become the most fundamental quantity describing a customer's delay or a queueing system's behavior. However, as Larson [9], So and Tang [14] and some other researchers reported (e.g., Schmitt *et al.* [13]), the customer's attitudes toward queues may be influenced more strongly by other factors, such as environment and social injustice. They have attempted to argue or demonstrate that at least three attributes other than the queueing waiting time play key roles in a customer's queueing experience: queueing environment, social justice and feedback about the delay. From their observations, customers usually feel better about queueing when they are provided with information that allows them to estimate in advance their waiting time in queue.

Many examples show that customer's queueing experiences and attitudes can impact a wide range of firms, including fast foods, department stores, banks and hotels, transportation services, emergency services, theme parks, airlines and so on. It seems clear that the environment in which queue waiting occurs plays a fundamental role in a customer's perceived and/or actual cost of participating in that system. The fraction of the total operating cost of a facility often can go a long way towards alleviating customer anguish and discomfort, perhaps even transforming it into well being and happiness. In terms of the management of waiting and delays, Schmitt *et al.* [13] suggest that service providers should not only be concerned about minimizing customers' average waiting time but also design queueing support systems that are susceptible to psychological and behavior factors.

A similar observation was published by Finlay *et al.* [8]. They considered the problem of organizing the medical residents' work in settings where queue demands are heavy and resources are limited. The focus of their work is framework of relationship among "multiple queues" resulting from the organization of patient care into a set of sequential activities in which patients are seen first by a nurse (one queue) and then by a resident (a second queue). The reason that a queue is divided into two queues in series may be explained as follows. Generally, patients in a clinic waiting in a guest room can occupy their time watching the magazines, newspapers or paintings. Waiting after a while, the patients are called by a

*Corresponding author: paul@math.nccu.edu.tw

pre-registered number assigned according to their arrival to enter a small waiting room in order to finish some paper work or pre-checking with nurses. Although these may not be necessary to complete before seeing a doctor, it makes the patients feel they are not longer in waiting, instead they are being served. However, no details of solution methods for this problem were addressed in evaluating its performance.

Summarizing our discussion, we have attempted to demonstrate that there exist instances in which the conventional notion, an average queue waiting time, does not reflect the effective measure properly. In this paper, we investigate some mediating factors that may be related to the environment as well as the structure of the queue itself. We show a semaphore queue modeling (SQM) can be used to analyze this setting design problem. Solution methods of this SQM under the environmental consideration are presented for the first time with theoretical reasoning.

In order to describe the fundamental factors that characterize the problem, we study the classic M/G/1 queueing system with a finite subqueue that represents a different waiting environment. Our goal is to construct such a queueing model and define its performance measures of effectiveness. As a consequence, we define new measurements to represent the delay and prove the modified queueing model is equivalent to a classic M/G/1 model but provides more information. Numerical examples are given to illustrate these results.

Several methods have been proposed to deal with finite M/G/1 queueing models, e.g., Carroll *et al.* [4] and Truslove [15]. Amongst the results obtained is the distribution of queue length at instants at which customers complete service. Our approach here differs from previous methods in many ways and it may be useful if we identify the contribution of this paper. First, we consider a queueing model with an infinite buffer but within it there is a finite queue to represent a different waiting environment. Second, in contrast to more complex analyses, the model we propose is rather simple and quickly understood.

The structure of this paper is organized in the following. We begin to introduce some performance measures that may describe the different environment in the next section. In Section 3, we present a simple M/G/1 queueing model with a finite subqueue. Its stationary probability distribution is derived and an analysis is made to compare models with different subqueue sizes in Section 4. In Section 5, we propose six different performance measures and then introduce an easy-implemented algorithm for an optimization problem. In Section 6, several numerical examples are illustrated to verify the results. Finally, we discuss implications of our results and possible extensions of the method.

2. OPPORTUNITY COST OF WAITING

2.1 Measuring Customers' Satisfaction

One important candidate for a measure of the psychological cost is the level of satisfaction which reflects the potential consequence of waiting for service. There are many factors that can contribute to a customer's satisfaction with the level of service received. Davis and Heineke [6] suggested that managers can influence the level of dissatisfaction with waiting times, to some extent, by managing customer perceptions of waiting times. In an attempt to improve the customers' perception of the waiting experience, Katz *et al* [10] in their empirical study of bank customers found that customer satisfaction was inversely related to customer perceptions of waiting time, e.g., an electronic news board which transmitted up-to-date news.

Having demonstrated the important factors that go beyond the exact waiting time considerations, the present studies focus on that individuals not only care about the objective length of a delay but also react to the environment in which such delays occur. Consider a classic M/G/1 queueing system where it adopts all conventional assumptions except its queue line is divided into two different waiting environments. Suppose these two subqueues are denoted by Q_1 and Q_2 in which Q_2 has a finite waiting space whose size is denoted by *N*, including one in service.

Let $q(u)$ be the dissatisfaction level at time u. Let $r_i(q)$ be the rate at which dissatisfaction changes during waiting in Q_i, i =1,2, when the level of dissatisfaction is *q*, namely,

$$r_i(q(u)) = \lim_{\tau \downarrow t} \frac{q(\tau) - q(u)}{\tau - u}$$

One sees that the level of dissatisfaction $q(u_1 + u_2)$ at time $u_1 + u_2$ is given by the following two equations depending on *u*:

$$q(u) = q(0) + \int_0^u r_1(q(x))dx \quad 0 \le u \le u_1$$

$$q(u) = q(u_1) + \int_{u_1}^u r_2(q(x))dx \quad u_1 \le u \le u_1 + u_2$$

where $q(0)$ represents the customer's initial dissatisfaction level, u_1 and u_2 denote times spent in Q_1 and Q_2, respectively. Let $h(u)$ be the level

of dissatisfaction at time t if service is not provided during [0, u] while $q(u)$ the level of dissatisfaction at time u not conditioned on service. Likewise, for any $u \geq 0$, we have

$$h(0) + \int_0^u \quad))dx \quad u \geq 0$$

It has been argued in Carmon *et al*. [3] that there exists $z > 0$ such that $q(u_1 + u_2) = h(u_1 + z)$ if $r_2 < 0$. When this happens, a so-called reduction in effective waiting time was discussed in their paper. Although the equality was not proved in their paper, it is mathematically correct if these functions are continuous and well-defined for its purpose.

However, it is not clear how this relates to the cost of waiting. As suggested by Davis [5], the cost of waiting can be written in terms of the opportunity costs associated with the average net dollars that is estimated by each customer's contribution which is expressed as a function of waiting time. Let $k(n)$ be a cost function of n customers waiting in the system in average. Based on these arguments, we would prove Lemma 1 by a similar approach taken in [3].

Lemma 1: If the unit cost per customer is proportional to the number of customers in system and $r_2(q)/r_1(q)$ is monotonic in q, then k(n) is monotonic in n.

Proof: Let $f(s,u)$ be the portion of the effective waiting time accumulated during the service period if the time of entering Q_2 is initiated when the level of dissatisfaction is $h(u)$ and is continued for s units of time. From the definition of $h(u)$, it observes that if $h(u) = x$, then $du/dx = 1/r_2(x)$ when the level of dissatisfaction is x. Since s is the duration of change in the level of dissatisfaction from $h(u)$ to $h(u+f(s,u))$, we have

$$s = \int_{h(u)}^{h(u+f(s,u))} \frac{}{r_2}$$

Taking the derivative of both sides of the equation above with respect to s, one gets

$$\frac{d}{ds} f(s,u) = \frac{r \quad + f(s,u))}{r \quad + f(s,u))}$$

Given a fixed duration s, it can be shown that $f(s,u)$ is monotonic in t as the right side increases or decreases. The detail of proof can be found in [3]. By Little's formula, for any subsystem A in a queue, the average number of customers in A equals to the average waiting time spent in A times the average arrival rate when the stability condition is assumed. Thus, it is apparent that $k(n)$ is proportional to $f(s,u)$.

Notice that $k(n)$ may be defined by the manager through situation dependent and customer-oriented data that reflect the particular relationship between a customer's level of satisfaction with waiting and his/her respective waiting time. However, it is still a function of a random variable, i.e., the number of customers in the system. To capture its characteristics, we give a prescriptive model and an algorithm for computation in the following sections.

2.2 Mean and Variance of the Delay

In all analysis, it is crucial to study the mediating factors that are pertinent to queueing systems. We investigate not only the average delay but also the variances of waiting times as the measures indicating the difference between two different environments. We propose six different effective measures including the classic measures of evaluating the performance of the queueing model, i.e., L, variance, W, ratio of W, ratio of variances, and a cost function (the last three are to be defined later). First two of them are the measures that reflect the mean and variance of lengths of the waiting lines in Q_1 and Q_2, say L_1, L_2, V_1 and V_2, respectively. In addition, because Q_2 is finite, a crucial measure is the probability of Q_2 being full.

Besides the traditional measures, we define LR and VR to be the natural logarithm of the ratio of L_2 and L_1, V_2 and V_1, respectively. Psychologically, customers review them as factors to examine the system's performance. The values of these ratios give customers waiting in Q_1 the perception of how queue behaves. The large value of LR means Q_2 is not fully utilized but the small value means customers feel congested in the system. On the other hand, VR reflects the comparison of variances of waiting lines between Q_1 and Q_2. Since both L_2 and L_1 are functions of N, the ratios are also functions of N. Neither too large nor too small, their values reveal how the system manager can plan with the size of Q_2, namely, N.

There are two reasons why we use ratios as the effective measures. First, the ratio is a function of N and well-behaved with respect to the traffic load. Second, while LR only characterizes the rational expression of linear functions VR may reflect the dispersion among queues. Furthermore, extending the notion of the queueing length, we are able to attain the best size for Q_2 so that the total cost is minimized subjecting to certain cost constraints. Consequently, using these measures, we will construct a queueing model of which the performance is evaluated

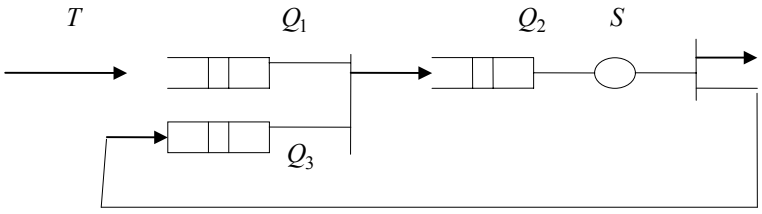

Figure 1. A M/G/1/(N) queueing mode

analytically.

3. THE SQM FORMULATION

In order to construct a subqueue in a queueing model, we use tokens to represent the customer in the subqueue and service conceptually. Suppose the number of tokens *N* is finite and is considered as the maximal capacity of the subqueue including one in service. Consider a single-server system shown in Figure 1. Customers arrive at Q_1 in accordance with a Poisson stream of rate λ will ask for the permission to enter the system. Only those customers who have tokens from queue Q_3 are legitimate to enter queue Q_2. Otherwise, they wait in Q_1 until tokens are available in Q_3. Those tokens in Q_2 shall not be returned to Q_3 until the customers finish the service in S. Such a token model has been widely-used to represent communication systems. One of them was proposed by Fdida *et al*. [7] to study the interaction and congestion on queues. In their model, there is more than one server and tokens are designed to control the data flow in communication networks. Their service times are assumed exponential and the method of Mean Value Analysis is taken. If the arrival process is assumed Coxian, then an approximation procedure is applied in [11]. However, the model we study has a general server, and the solution derived is exact while we make use of tokens to represent customers in the subqueue.

The system described above is denoted by M/G/1/(*N*) where (*N*) represents a finite subqueue with size of *N*. Our intent here is to valid this closed queueing model of tokens and use this model to solve the setting design problem. Assume there is no delay for returning a token to Q_3. The service discipline is first-come-first-served. The server can only serve one customer each time and its service time distribution is general and independent of the arrival process. The size of Q_1 is unlimited. Statistically, service times are assumed to have a general distribution function *H*(t), $t \geq 0$, with $H(0^+) < 1$ where *H*(.) has a finite expectation $1/\mu$. The set of all service times is assumed a stochastically independent family. For a stable system, we assume $\mu > \lambda$.

Let $(\tau_n : n = 0,1,2,...)$ be the sequence of departure times where $\tau_0 = 0$. Consider the process $\Lambda = \{(c_n, t_n) : n = 0,1,2,...\}$ embedded at $(\tau_n : n = 0,1,2,\cdots)$ where c_n and t_n are the number of customers and tokens presenting at Q_1 and Q_3, respectively; then the process Λ is a Markov chain on the state space $U \triangleq \{0,1,2, \} \times \{0,1,2,..N\}$. Under the condition of stationary for a state (*c*, *t*), we mean there are *c* customers in Q_1, *t* tokens in Q_3. Because the arriving customer must go into the system if there is one token in Q_3; or, it is accumulated in Q_1 when there is no token, we have $c \cdot t = 0$ at any time.

According to this embedded Markov chain in Λ, transitions in or out of the state (*c, t*) can be caused by the occurrence of one of the following events:

1. Transition from state (*0*, *N*), i.e., the state where no customer in service after a departure.
2. Transition from state (c+1, *t*) as *t* = 0, i.e., a token returns to Q_3 and goes to Q_2 immediately with a customer and no arrival during the latest service.
3. Transition to state (*c*, *t*+1) as *c* = 0, i.e., a token returns and waits in Q_3, and no customer arrive at Q_1 during the latest service.
4. Transition to state (*c*+ *i*, *t*) as *t* = 0, i.e., (*i*+1) customers arrive at the system during the latest service and a token returns to Q_3 and goes to Q_2 immediately with a customer.
5. Transition to state (*i*-*t*,0) as *c*=0, i.e., *i*+1 customers arrive in the system during the latest, for $i \geq t$.
6. Transition to state (c, *t*-*i*) as *c*=0, i.e., *i*+1 customers arrive in the system during the latest service, for $t \geq i$.

Let A_i be the arrival time instant for the *i*th customer entering the system, *i* =0,1,2,… where $A_0 = 0$. Let $T_i = A_i - A_{i-1}$ be the interarrival time between the (*I*-1)th customer and the *i*th customer for *i*= 1,2,…. By the assumption we adopt, $T_1, T_2, .$ are independent and identically exponential distributed. Let *Y* be a random variable distributed according to

H (.). Let δ_v and σ be the probability having more than *v* arrivals and no arrivals during the service time *Y*, respectively, i.e.,

$$\delta_v = \Pr\left\{ \sum_{i=1}^{v} T_i < Y \right\}$$

$$\sigma = \Pr\{ T > Y \}$$

Lemma 2: Consider all states in U, the probabilities of their transitions are given below:

(a) For $0 \le t \le N$, $c \ge 0$
$\Pr((0, N), (c,t)) = \Pr((0, N-1), (c,t))$

(b) For $0 \le t \le N-1$
$\Pr((0, t), (0, t+1)) = \sigma$

(c) For $i \ge 0$, n > 0
$\Pr((c,0), (c+i,0)) = \delta_{i+1} - \delta_{i+2}$

(d) For $i \ge 0$
$\Pr((i+1,0), (i,0)) = \sigma$.

(e) For $i > t$, $0 \le t < N$
$\Pr((0, t), (i-t,0)) = \delta_{i+1} - \delta_{i+2}$

(f) For $0 \le i \le t$, $0 \le t < N$
$\Pr((0, t), (0, t-i)) = \delta_{i+1} - \delta_{i+2}$

Proof: Part (b) and (d) are the probabilities that no customers arrive during the service time *Y*, that is $\Pr\{T>Y\}$. Part (c), (e), and (f) are the probabilities that exactly *i*+1 customers arrive during the service time *Y*. Part (a) is expressed for the case of no customers in the service after a departure occurs. The probability of transition from this case to any system state is equivalent to that of transition from exactly one customer left in the system after a departure to any system state. This is because the Markov chain is considered embedded at the departure instants.

4. TRANSITION PROBABILITY MATRIX AND STATE BALANCE EQUATIONS

From Lemma 2, we observe that

$$\Pr\{(0,k),(0,k+1)\} = \Pr\{(i+1,0),(i,0)\}$$

and

$$\Pr\{(n,0),(n+i,0)\} = \Pr\{(0,k),(i-k,0)\}$$
$$= \Pr\{(0,k),(0,k-i)\}$$

Though they represent for different conditions on the state transitions, to be concise in presentation, the following elements are defined in the transition matrix. Let

$$b_0 \triangleq \Pr\{(i+1,0),(i,0)\} \quad i \ge 0$$
$$b_{i+1} \triangleq \Pr\{(n,0),(n+i,0)\}$$
$$i \ge 0 \quad \text{for any fixed } n$$

Let $P_{(N)}$ be the transition matrix, given *N*, of the Markov chain underlying the process described by the state (*c*, *t*) that has an equilibrium transition probability

$$\Pr\{(c_1,t_1),(c_2,t_2) \mid (c_1,t_1),(c_2,t_2) \in \mathrm{U}\}$$

$P_{(N)}$ is thus expressed by

$$P_{(N)} = \begin{bmatrix}
b_1 & b_0 & & & & & b_2 & b_3 & \cdots & \cdots \\
b_2 & b_1 & b_0 & & & & b_3 & b_4 & \cdots & \cdots \\
b_3 & b_2 & b_1 & b_0 & & & b_4 & b_5 & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \cdots & \cdots & & & \\
b_N & b_{N-1} & \cdots & & b_1 & b_0 & b_{N+1} & b_{N+2} & \cdots & \cdots \\
b_N & b_{N-1} & \cdots & & b_1 & b_0 & b_{N+1} & b_{N+2} & \cdots & \cdots \\
b_0 & & & & & & b_1 & b_2 & b_3 & \cdots \\
 & & & & & & b_0 & b_1 & b_2 & \cdots \\
 & & & & & & & b_0 & b_1 & \cdots \\
 & & & & & & & & b_0 & \cdots \\
 & & & & & & & & & \cdots
\end{bmatrix}$$

In sequel, we shall discuss how to obtain the steady state probability of the Markov chain. Let π be the steady state probability vector of Markov generator underlying this queue, where π_i denote the probability of having *i* tokens in Q_3 for $0 \le i \le N$, π_{N+i} denote the probability of having *i* customers in Q_1 for $i \ge 1$, and π=($\pi_0, \pi_1, ..., \pi_N$, $\pi_0, \pi_1, ..., \pi_N, \pi_{N+1}, ...$).

By the stationary assumption, the vector π satisfies the following equation:

$$\pi \cdot P_{(N)} = \pi$$

The equilibrium equations for this process are

$$\pi_0 b_1 + \pi_1 b_2 + \cdots + \pi_{N-1} b_N + \pi_N b_N + \pi_{N+1} b_0 = \pi_0 \tag{1}$$

$$\sum_{j=1}^{N} \pi_{j-1} b_{j-i} + \pi_N b_{N-i} = \pi_i \tag{2}$$
$$i = 1, 2, \cdots, N-1$$

$$\pi_{N-1} b_0 + \pi_N b_0 = \pi_N \tag{3}$$

$$\sum_{j=0}^{N-1}\pi_j b_{j+k+1} + \pi_N b_{N+k} + \sum_{j=0}^{k}\pi_{N+k+1-j} b_j = \pi_{N+k} \quad k = 1, 2, \cdots \tag{4}$$

$$\sum_{i=0}^{\infty}\pi_i = 1 \tag{5}$$

In our approach, it does not need solving any system of equations because all π_i can be expressed by π_N only which is the probability of N tokens in Q_3.

Rearranging the equations (3) (2) and (1), we have

$$\pi_{N-1} = \pi_N(\frac{1}{b_0} - 1) \tag{6}$$

$$\pi_{i-1} = \left[\pi_i - \sum_{j=i+1}^{N}\pi_{j-1} b_{j-i} - \pi_N b_{N-i}\right]\cdot(1/b_0) \tag{7}$$

$$i = N\text{-}1, N\text{-}2, \ldots 1,$$

$$\pi_{N+1} = [\pi_0(1-b_1) - \sum_{j=1}^{N-1}\pi_j b_{j+1} - \pi_N b_N]\cdot(1/b_0) \tag{8}$$

To solve π_i, for $i > N+1$, it is clear from (4) to write π_{k+N+1} in termof π_n, $n = 1, 2, \ldots, N+1$, for all $k \geq 1$, that

$$\pi_{N+k+1} = \left[\pi_{N+k} - \sum_{j=0}^{N-1}\pi_j b_{j+k+1} - \pi_N b_{N+k} - \sum_{j=1}^{k}\pi_{N+k+1-j} b_j\right] . (1/b_0) \tag{9}$$

Lemma 3: The steady state probability of server being idle is π_N.

Proof: At any point τ_n, if $t_n = N$, then we have $c_n = 0$. This means there is no customer in the system at the departure point τ_n. Under the assumption we adopt, $\lambda < \mu$, the system is stable and the limit exists at the departure point. Also by the theorem of PASTA (Poisson Arrivals See Time Average) in Wolff [16], we know the system is ergordic and the limit at any time is equivalent to the limit at the departure instant. Explicitly, π_N is the steady state probability of server being idle at any time. Let the traffic load be $\rho \triangleq \lambda/\mu$. Since the probability of the system being idle in steady state is $(1-\rho)$, we have $\pi_N = 1-\rho$.

By Lemma 3, the steady state probabilities given by (6), (7), (8) and (9) are easy to compute. Starting with $1-\rho$, all π_i can be substituted and obtained recursively one by one. Because of $b_0 > 0$, the steady state probability distribution is uniquely determined by (6), (7) (8) and (9).

Theorem 1: Under assumption of $\lambda < \mu$, the steady state probability within Λ is solved by (6), (7) (8) and (9).

From Lemmas 2 and 3, the proof is straightforward and omitted here.

Denote by $\pi_i(N)$ the steady-state probability when the buffer size of Q_2 is N with respect to (N-i) customers in system if $i \leq N$ or i customers in system if $i > N$. The following corollary describes the probability distributions relating different sizes of Q_2.

Corollary 1: Consider two queueing models, M/G/1/(N) and M/G/1/(K) with the same interarrival and service time distributions. The queue sizes of Q_2 are N and K, respectively. Suppose N > K. Then we have

(a) $\pi_{N-K+i}(N) = \pi_i(K),\ i = 0,1,2,\ldots,K$

(b) $\pi_{N-K-i}(N) = \pi_{K+i}(K),\ i = 1,2,\ldots,N\text{-}K$

(c) $\pi_{N+i}(N) = \pi_{N+i}(K),\ i = 1,2,\ldots$

The proofs are given in the appendix.

From the definition of W, it can be written as

$$W = \frac{1}{\lambda}[\sum_{n=0}^{N}(N-n)\cdot\pi_n + \sum_{n=1}^{\infty}(N+n)\cdot\pi_{N+n}]$$

Let $W(N)$ and $W(K)$ be the mean waiting time computed by N and K, respectively. We compare the elements in computing their mean waiting time. It yields

$$W(N) = \frac{1}{\lambda}[N\cdot\pi_0(N) + (N-1)\cdot\pi_1(N) + \cdots + (K+1)\cdot\pi_{N-K-1}(N) + \sum_{i=0}^{K}(K-i)\cdot\pi_{N-K+i}(N) + \sum_{i=1}^{\infty}(N+i)\cdot\pi_{N+i}(N)]$$

and $W(K) = \frac{1}{\lambda}[\sum_{i=0}^{K}(K-i)\cdot\pi_i(K) + \sum_{i=1}^{\infty}(K+i)\cdot\pi_{K+i}(K)]$

From Corollary 1, we have $W(N) = W(K)$

Let a_n be the steady state probability of n customers in M/G/1, e.g., $a_0 = \pi_N$, $a_n = \pi_{N-n}$ for $1 \leq n \leq N-1$, and $a_n = \pi_n$ for $n > N$. Notice N only defines various subqueue configurations which just have different permutations of a

probability vector. Using the fact that the steady state probability is permutation-invariant, we have the following theorem.

Theorem 2: The M/G/1/(N) model for any positive integer N is equivalent to a classic M/G/1 model.

This theorem can be immediately obtained since all of these models have the same steady state probability distribution.

Given the size of Q_2, in terms of a number of tokens available, Procedure 1 in the following determines the stationary probabilities.

Procedure 1: A solution procedure to the M/G/1/(N)

Step 1: Let $\pi_N = 1-\rho$.

Step 2: Obtain π_i for i = 1...N-1.

Step 3: Obtain π_i for i> N.

5. PERFORMANCE MEASURES

In this section, our intention is to provide computing formulas to calculate the required performance measures defined in Section 2. Given π obtained from Procedure 1, the probability of Q_2 being full is expressed in terms of no token in Q_3. The probability of no token in Q_3, π^*, is given by

$$\pi_0^* = \pi_0 + \sum_{n=1}^{\infty} \pi_{N+n}$$

The average queue length L_1 is computed by

$$L_1 = \{1 \cdot \pi_{N+1} + 2 \cdot \pi_{N+2} + \cdots + n \cdot \pi_{N+n} + \cdots\}$$

The mean number of customers waiting in Q_2 and in service is given by

$$L_2 = \sum_{\mathrm{n}=1}^{N} (N - \mathrm{n}) \cdot \pi_{\mathrm{n}} + N \cdot \pi_0^*$$

According to Little's formula, the mean waiting time in this model is also easily computed. Similarly, we calculate the variances by definition at Q_1 and Q_2, known as V_1 and V_2:

$$V_1 = \{1^2 \cdot \pi_{N+1} + 2^2 \cdot \pi_{N+2} + \cdots + n^2 \cdot \pi_{N+n} + \cdots\} - L_1^2$$

$$V_2 = \sum_{\mathrm{n}=1}^{N} (N - \mathrm{n})^2 \cdot \pi_{\mathrm{n}} + N^2 \cdot \pi_0^* - {L_2}^2$$

Thus, LR and VR as defined in Section 2 are computed respectively by

$$\mathrm{LR}(N) = \ln(\frac{L_2}{L_1})$$

$$\mathrm{VR}(N) = \ln(\frac{V_2}{V_1})$$

where ln(.) is a natural logarithm.

Since L_1 and L_2 are functions of N and L_1 decreases, L_2 increases as N increases, we have $\mathrm{LR}(N)$ increases as N increases. Similarly, because V_1 decreases and V_2 increases as N increases, it results in $\mathrm{VR}(N)$ increases as N increases. Thus, we present the following two lemmas.

Lemma 4: For fixed λ and μ, VR(N) and LR(N) are nondecreasing functions of N.

The proof may be derived from the description above.

Lemma 5: For a fixed N, VR(N) and LR(N) are both functions of ρ over (0,1) and are nonincreasing on this interval.

Proof: First, we consider *LR(N)*. Because N is fixed and L_2 is the mean queue size of a finite queue but L_1 is computed for an infinite queue, the ratio of them is nonincreasing as the traffic load approaches to 1. Therefore, *LR(N)* is nonincreasing as well. Then, the similar arguments are applied to *VR(N)*..

Let LR(N) and LR(K) be the ratios of the mean waiting time calculated with respect to N and K. Similarly, denote VR(N) and VR(K) be the ratios of variances, respectively.

Theorem 3: Given N and K, |VR(N)-VR(K)| or |LR(N)-LR(K)| is a function of ρ over (0,1) and is nonincreasing on this interval.

Proof: To avoid a trivial case, assume $N > K$. First, we consider VR(N)-VR(K). By Lemma 4, we know VR(N)>VR(K) for all $N > K$. By Corollary 1, we have $\pi_i(N) = \pi_i(K)$ for all i as both N and K approaches to infinity. It implies

$$\lim_{N\to\infty} \frac{1}{\mathrm{VR(N)}} = \lim_{K\to\infty} \frac{1}{\mathrm{VR(K)}}$$

Hence, |VR(N)-VR(K)| is nonincreasing and its lower bound is 0. A similar proof may be applied to |*LR(N)-LR(K)*|.

In the remainder of this section, we discuss for a minimization problem regarding two comparable environments. Extending the notion of the queueing length, we consider a cost function associated with the number of customers in Q_2, for example,

customers perceive a minute of delay in Q_2 may be two or three times that in Q_1. Suppose the cost function of having n customers in Q_2 is $k(n)$, as defined in Section 2. The average total cost $J(N)$ is defined by

$$J(N) = \sum_{n=1}^{N} k(N-n) \cdot \pi_n + k(N) \cdot \pi_0^*$$

If $k(n)$ is an identity function, i.e. $k(n) = n$, then

$$J(N) = \sum_{n=1}^{N} (N-n) \cdot \pi_n + N \cdot \pi_0^* = L_2$$

Under the assumption of Lemma 1, if $k(n)$ is convex, $J(N)$ is not necessarily convex since N is a positive integer implying $J(N)$ is not continuous. However, $J(N)$ is unimodal which would be proved in the following theorem.

Theorem 4: Given a convex function ,k(n) over [a, c], where a and c are integers and chosen with managerial choice regarding customers' satisfaction level, J(N) is unimodal on [a, c] for $a \le N \le c$.

Proof: Without loss of generality, assume there exists an N* at which J attains a minimum. By definition of unimodal, we shall claim J is nondecreasing for $N \ge N$ * and nonincreasing for $N \le N$ *. Let $\nabla J(N+1) = J(N+1) - J(N)$. Thus , we have

$$\nabla J(N+1) = \sum_{n=1}^{N} \{k(N+1-n) - k(N-n)\} \cdot \pi_n + \{k(N+1) - k(N)\} \cdot \pi_0^* + k(0) \cdot \pi_{N+1}$$

Consider two cases below with the fact that $k(n)$ is monotonic which is proved in Lemma 1.

Case 1: $N \ge N$ *. Assume $\nabla J(N+1) < 0$. Since $f(x)$ is convex and all $\pi_i > 0$ for all i, it leads to a consequence that N * does not minimize $J(N)$. It contradicts to our assumption. Therefore, we have $\nabla J(N+1) \ge 0$.

Case 2: $N < N$ *. Because N is an integer, we further assume $N+1 \le N$ *. The proof will follow the similar arguments taken above. In contrast to proving $\nabla J(N+1) \ge 0$, we prove $\nabla J(N+1) \le 0$ in this case. However, this is omitted for its similarity. Hence, we have checked J is nondecreasing for $N \ge N$ * and nonincreasing for $N \le N$ * which completes the proof.

Clearly, minimizing J is a nonlinear integer programming problem. It is worth mentioning that among the derivative-free methods that minimize unimodal functions over a closed bounded interval, the Fibonacci search method is the most efficient in that it requires the smallest number of observations for a given reduction in the length of the interval of uncertainty. The details of the solution appear in Bazaraa *et al.* [1]. Nevertheless, notice that minimizing J while solving Procedure 1 for each configuration of possible N is time-consuming because every configuration has different permutations of a probability vector. Instead, using the fact that π_N is permutation-invariant, i.e., $\pi_N = 1 - \rho$, we propose a heuristic method which minimizes J by taking Corollary 1 into account. Moreover, because N * minimizing J may not be an integer, Procedure 2 that searches for an integer $\overline{N}$ which is suboptimal to N * is described as below.

Procedure 2: Search for $\overline{N}$

Step0: Let $N_1 = a$.

Step1: If $\nabla J(N_1 + 1) \ge 0$ then go to step3;
else let $N_1 = N_1 + 1$ and go to step2.

Step2: If $N_1 \ge c$, then go to step3;
else go to step1.

Step3: Print $\overline{N} = N_1$ and stop.

Given the traffic load, the Procedure 2 is to determine the suboptimal number of tokens, although the objective is to minimize the cost function. The proposed method can iterate over the different number of tokens starting with the smallest number and incrementing by one until it equals a suboptimal point. Since in practice a domain of continuous values of N is not permitted, the resulting value serves as a lower bound to the optimal value of the original problem.

Solving the original problem where minimizing the average total cost is replaced by the decision of allocating the number of tokens to the system. The replacement allows a facilitation in the setting design problem. Furthermore, since π provides target performance as part of the solution, the rationale behind this model is that π_N remains the same for different capacities at Q_2 but $J(N)$ is unimodal and well behaved with respect to the subqueue size.

6. NUMERICAL ILLUSTRATE EXAMPLES

In this section, we present numerical experiments describing the various aspects of system behavior for different values of arrival rate λ with rate fixed to 1. In this case, we have $\rho = \lambda$ while studying the system's behavior.

Consider three different service distributions in examples whose probability density functions are, Erlang-2, E_2, hyper-exponential, H_2 and Erlang-4,

E_4, respectively,

$$\text{E\_2}\ (t)=4te^{-2t},\ t\geq 0$$

$$\text{H\_2}\ (t)=0.5\ \ e^{-1.5t}+0.5\,e^{-0.75t},\ t\geq 0$$

$$\text{E\_4}\ (t)=\frac{2}{3}(4t)^{3}\ \ e^{-4t},\ t\geq 0$$

Each example is tested with two different buffer sizes 3 and 5. All test problems are computed with $\mu=1$, and λ is taken from 0.1 to 0.9. From the illustrative examples, the mean waiting time are studied and found in the following order: M/ H_2/1/(3), M/E_2/1/(3), M/E_4/1/(3). The M/ H_2/1/(3) queueing model gives the highest waiting time. This is demonstrated in Figure 2.

Conversely, M/ E_4/1/(3) gives the highest VR and M/H_2/1/(3) gives the lowest VR but they all are lower than that of 5-token models. This is shown in Figure 3. In Figure 4, it shows LR versus traffic load with respect to M/E_4/1/(3) and M/ E_4/1/(5).

Clearly, it shows LR increases as N increases and decreases as the traffic load approaches to 0.9. Figure 4 shows it is almost impossible to have an equal queue length between Q_1 and Q_2, i.e., LR=0, unless the traffic is very heavy. However, Figure 3 shows it is not difficult to let two queues with equal variance, i.e., VR=0, within a moderate traffic. Notice customers in light traffic are not sensitive to the environment since most of them do not spend time in queue. The closer the ratios obtained from the traffic load to 1, the less different are the two environments. Therefore, these measures are suitable for cases in moderate traffic.

7. CONCLUSIONS

For the delay been considered, the variance is computed from the SQM as function of the number of tokens and the traffic load. Given available arrivals and their service requirements, the problem of setting design is to simultaneously find the subqueue sizes or assign the number of operational tokens. The objective here is to characterize VR and LR, the system's behavior in terms of the ratio of mean waiting time and their variances between two queues. It turns out both VR and LR have an interesting characteristic that suggests the system manager a viewpoint different from traditional notion and may reflect an effective measurement.

We use a closed queueing network model to represent the subsystem. Our next research problems may be the study of characterizing LR and VR in more complicated cases. For example, consider the city bus services. Passengers have to experience two stages of waiting before reaching their destination: one is at a bus stop; the other is in the bus. Passengers usually feel better about queueing when they are in the bus rather than at the bus stop. Since the bus' capacity is finite, an interesting problem is to estimate its optimal capacity while satisfying certain cost constraints. This challenging problem may be considered as an example of a closed queueing network model with population constraints.

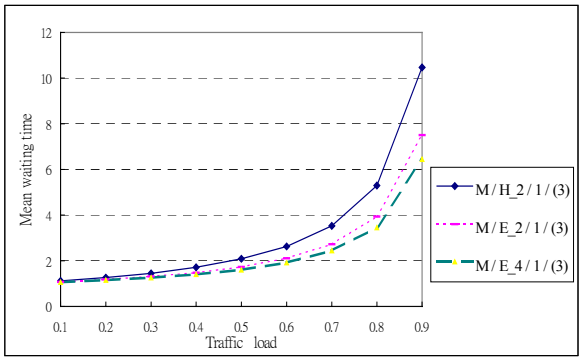

Figure 2. Mean waiting time vs. Traffic loads w.r.t. M/ E_2/1/ (3), M/ H_2/1/ (3) and M/ E_4/1/ (3)

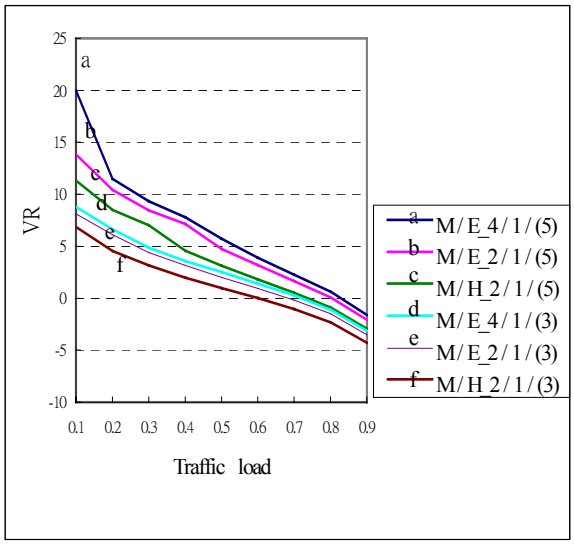

Figure 3. LR vs. Traffic loads w.r.t. M/E_2/1/(3), M/H_2/1(3) and M/E_4/1/(3)

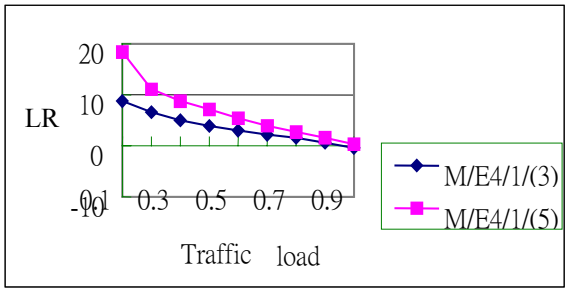

Figure 4. LR vs. Traffic load w.r.t. M/E_4/1(3) and M/ E_4/1/(5)

APPENDIX

The proof for Corollary 1:
(a) If the queue size of Q_2 is N, then from (6) we have

$$\pi_{N-1}(N) = \pi_N(N)(\frac{1}{b_0} - 1)$$

and from (7) we have

$$\pi_{N-2}(N) = (\pi_{N-1}(N) - \pi_{N-1}(N) \cdot b_1 - \pi_N(N) \cdot b_1)(1/b_0)$$

Now, if the queue size of Q_2 is N-1, then from (6) again we have $\pi_{N-2}(N-1) = \pi_{N-1}(N-1)(\frac{1}{b_0} - 1)$

and from (7) we have

$$\pi_{N-3}(N-1) = (\pi_{N-2}(N-1) - \pi_{N-2}(N-1) \cdot b_1 - \pi_{N-1}(N-1) \cdot b_1)(1/b_0)$$

From the fact that

$$\pi_N(N) = \pi_{N-1}(N-1) = 1 - \lambda/\mu$$
$$\pi_{N-1}(N) = \pi_{N-2}(N-1)$$

and by induction we have

$$\pi_{N-K+i}(N) = \pi_i(K), \ i = 0,1,2,...,K$$

(b) From (8), we have

$$\pi_{K+1}(K) = [\pi_0(K) - \pi_0(K) \cdot b_1 - \sum_{j=1}^{K-1} \pi_j(K) \cdot b_{j+1} - \pi_K(K) \cdot b_K] \cdot (1/b_0)$$

and from (7)

$$\pi_{N-K-1}(N) = [\pi_{N-K}(N) - \sum_{j=N-K}^{N-1} \pi_j(N) \cdot b_{j-N+K} - \pi_N(N) \cdot b_K] \cdot (1/b_0)$$

So, it implies $\pi_{K+1}(K) = \pi_{N-K-1}(N)$
Again, from (9), we have

$$\pi_{K+2}(K) = [\pi_{K+1}(K) - \sum_{j=0}^{K-1} \pi_j(K) \cdot b_{j+2} - \pi_K(K) \cdot b_{K+1} - \pi_{K+1}(K) \cdot b_1] \cdot (1/b_0)$$

and from (7) we have

$$\pi_{N-K-2}(N) = [\pi_{N-K-1}(N) - \sum_{j=N-K}^{N} \pi_{j-1}(N) \cdot b_{j-(N-K-1)} - \pi_N(N) \cdot b_{K+1}] \cdot (1/b_0)$$

It observes $\pi_{K+2}(K) = \pi_{N-K-2}(N)$
Consequently, by induction, we get

$$\pi_{K+i}(K) = \pi_{N-K-i}(N) \ \ i = 1,2,..., N\text{-}K$$

(c) From (1), (2) and (8), we have

$$\pi_{N+1}(N) = [\pi_0(N)(1 - b_1) - \sum_{j=1}^{N-1} \pi_j(N) \cdot b_{j+1} - \pi_N(N) \cdot b_N] \cdot (1/b_0)$$

and from (9), we have

$$\pi_{N+1}(K) = [\pi_N(K) - \sum_{j=0}^{K-1} \pi_j(N) \cdot b_{j+N-K} - \pi_M(K) \cdot b_N - \sum_{j=1}^{N-K} \pi_{N+1-j} \cdot b_j] \cdot (1/b_0)$$

It results $\pi_{N+1}(N) = \pi_{N+1}(M)$
Hence, by induction, we get

$$\pi_{N+i}(N) = \pi_{N+i}(K) \ \ i = 1,2,...$$

The proof is completed.

ACKNOWLEDGMENT

This research work was partially supported by the project under National Science Council, No. NSC 90-2218-E-004-002.

REFERENCES

1. Bazzraa, M. S., H. D. Sherali and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, New York, NY (1993).
2. Buzacott, J. and J. G. Shanthikumar, *Stochastical Models of Manufacturing Systems*, Prentice-Hall, Upper Saddle River, NJ (1993).
3. Carmon, Z., J. G. Shanthikumar, and T. F. Carmon "A psychological perspective on service segmentation models: the significance of accounting for customers' perceptions of waiting and service," *Management Science*, **41(11)**, 1806-1815 (1995)
4. Carroll, J. L., A. D. E. Liefvoorf and L. Lipsky, "Solutions of M/G/1//N-type loops with extensions to M/G/1 and GI/M/1 queues," *Operations Research*,

30(3), 490-514 (1982).

5. Davis, M. M., "How long should a customer wait for a service?," *Decision Sciences*, **22(2)**, 421-434 (1991).
6. Davis, M. M. and J. Heineke, "How disconfirmation, perception and actual waiting times impact customer satisfaction," *International Journal of Service Industrial Management*, **9(1)**, 64-73 (1998).
7. Fdida, S., H. G. Perros and A. Wilk, "Semaphore queues: modeling multilayered window flow control mechanisms," *IEEE Transactions on Communications*, **38**, 309-317 (1990).
8. Finlay, W., E. J. Mutran, R. Zeitler and C. Randall, "Queues and care: how medical residents organize their work in a busy clinic," *Journal of Health and Social Behavior*, **31**, 292-305 (1990).
9. Larson, R. C., "Perspectives on queues: social justice and the psychology of queueing," *Operations Research*, **35(6)**, 895-905 (1987).
10. Katz, K. L., B. L. Larson and R. C. Larson, "Prescription for the waiting-in-line blues: entertain, enlighten, and engage," *Sloan Management Review*, **32(2)**, 44-53 (1991).
11. Luh, H., "An approximation Procedure for multi-layered semaphore queues with Coxian arrivals and exponential service times," *International Journal of Computers and Their Applications*, **4(1)**, 47-55 (1997).
12. Robertazzi, T. G., *Computer Networks and Systems: Queueing Theory and Performance Evaluation*, Springer-Verlag, New York, NY (1990).
13. Schmitt, B. H., L. Dube, and F. Leclerc, "Intrusions into waiting lines: does the queue constitute a social system," *Journal of Personality and Social Psychology*, **63(5)**, 806-815 (1992).
14. So, K. C. and C. S. Tang, "On managing oper-ating capacity to reduce congestion in service systems," *European Journal of Operational Research*, **92**, 83-89 (1996).
15. Truslove, A. L., "Queue length for the M/G/1 queue with finite waiting room," *Advanced Applied Probability*, **7**, 215-226 (1975).
16. Wolff, R. W., *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Upper Saddle River, NJ (1989).

ABOUT THE AUTHOR

Hsing Luh received the Ph.D. degree in Operations Research from North Carolina State University, USA. He is Professor at the Department of Mathematical Sciences in National Chengchi University, Taiwan. He is interested in Simulation, Mathematical Programming, Queueing Networks and related problems.

(Received November 2001; revised February 2002; accepted April 2002)

排隊等候的另一種評估方法

陸 行*

國立政治大學應用數學系

116 台北市文山區指南路二段64號

摘要

等候理論的研究中，分析有關延遲等待的可能性已有數十年。在一個等候系統中，所有的平均等候時間及平均在系統內的時間已成爲觀察這個等候系統的一個基本指標。很明顯地，等候環境對顧客而言扮演一個很重要的角色。例如，改變子等候空間的環境讓顧客能減少在等候時的痛苦或不舒服。這個有趣的問題引發了我們去建立一個在不同觀察點的等候系統來修正等候模型的動機。在這論文中，我們將研究服務速率、顧客數目的變化與等候環境有關的等候模型。

關鍵詞：等候理論，等候成本

(*連絡人: paul@math.nccu.edu.tw)