Quality Technology & Quantitative Management
Vol. 2, No. 1, pp. 109-121, 2005

QTQM

Optimal Capacities of Tokens at Tandem-Queue Models of General Service Times

Hsing Luh and Chun-Lian Huang

Department of Mathematical Sciences, National ChengChi University, Taiwan
(*Received* May 2003, *accepted* August 2004)

Abstract: We consider a queueing system with two stations in series. Assume the service time distributions are general at one station and a finite mixture of Erlang distributions at the other. Exogenous customers should snatch tokens at a token buffer of finite capacity in order to enter the system. Customers are lost if there are no tokens available in the token buffer while they arrive. To obtain the stationary probability distribution of number of customers in the system, we construct an embedded Markov chain at the departure times. The solution is solved analytically and its analysis is extended to semi-Markovian representation of performance measures in queueing networks. A formula of the loss probability is derived to describe the probability of an arriving customer who finds no token in the token buffer, by which the throughput and the optimal number of tokens are also studied.

Keywords: Embedded Markov chains, probability distributions, queueing networks.

1. Introduction

In this paper we consider an open queueing model of finite capacity. The system consists of two stations in series and a token buffer whose size is limited. The simple structure of the model is depicted in Figure 1 and explained as follows. There are only one server and one queue at each station. Customers arrive from the exterior of the system following a Poisson distribution, entering the system at station 1 and departing at station 2. We assume the service time follows a finite mixture of Erlang distributions at serve 1 but a general distribution at the serve 2. The service discipline is First-Come-First-Served.

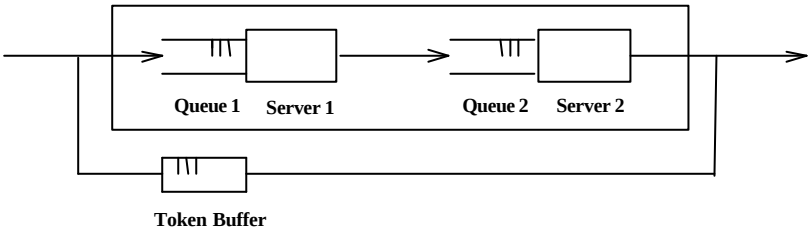

Figure 1.

Suppose the size of the token buffer is fixed and denoted by N. An arrival may enter station 1 if there is at least one token available at the token buffer, or it will be rejected by the system, thus resulting in a lost customer. Customers attend stations in sequence to obtain two services provided by server 1 and server 2. They leave the system and return tokens to the token buffer immediately when they complete their works. When a customer takes a token he will join server 1 if the server is idle, but wait for service in queue 1 if server 1 is busy. After

finishing service in server 1, he then joins server 2 if the server is idle otherwise he waits in queue 2. Each server can serve only one customer at a time. Assume either the capacity of queue 1 or queue 2 is at least as big as N. Hence, there is no balking of service at station 1 if server 2 is busy. Such a fundamental model can be used for studying the performance on a manufacturing line in which tokens may represent workers or kanbans while the customers stands for jobs. Buzacott and Shanthikumar [1] have pointed out the number of tokens is an important decision that affects jobs' travel time and throughput. For instance, too many tokens can lead to a poor flow time in the system, causing too much cost. To attain analysis of the behavior of N, both the joint stationary probabilities and the average number of tokens waiting in the token buffer are studied.

If there is only one server in the network, then the model is the same as the $M/G/1/N$ loss system where N is the maximum number of customers in the system. Miller [4] has derived a recursive formula of Laplace-Stieltjes transform for the mean value of the distribution of the lengths of busy periods for the $M/G/1$ finite queue. If there is no external arrival, and N tokens are replaced by N jobs in the network, the model becomes a closed model which has been considered in Daduna [2]. He derived the Laplace-Stieltjes transform of the job's cycle time, given the state of system at the beginning of the cycle. A modification of it was made to calculate the departure process by Luh [3]. Based on it, we will calculate a probability that an arriving customer finds no token in the token buffer, which is called a loss probability. Since the loss probability is a function of the number of tokens, of interest here is the study of a loss probability subject to a prescribed level of throughput. It will be carefully analyzed and developed in this paper according to the property of the joint stationary probability.

The paper is organized as follows. In Section 2, we give a detailed description of this model and construct an embedded Markov chain. In Section 3, by considering the embedded Markov chain we obtain the stationary distribution at the system. In Section 4, we present an analysis to compute the loss probability. In Section 5, a procedure of choosing an optimal number of tokens that minimizes the expected cost is also developed. Numerical examples are given for illustrating the property of the loss probability in terms of the number of tokens in the system.

2. Model Description

Consider a queueing system with Poisson arrivals, which consists of a network including two stations in tandem, single-server at each station and a token buffer as introduced in Section 1. Assume that the total number of the tokens available in the network is fixed to a maximal capacity N. Equivalently, the maximal capacity of each station is no less than N.

In addition, we have the following assumptions. The customers arrive according to a Poisson distribution with parameter l, and the service times at server 2 have a general distribution function $F(t)$, for $t \geq 0$, with $F(0^+) < 1$, and has a finite expectation. The service times at server 1 are taken from a finite mixture of Erlang distributions with phase parameter M, in which the service time at each phase follows an exponential distribution with parameter l'. In specific, the finite mixture of Erlang distributions is written as follows,

$$G(t) = \sum_{g=1}^{M} q_g \cdot \Gamma_{l',g}(t) = \sum_{g=1}^{M} q_g \cdot \left(1 - \sum_{h=0}^{g-1} e^{-l't} \cdot \frac{(l't)^h}{h!} \right),$$

$$\sum_{g=1}^{M} q_g = 1, \; q_M > 0.$$

Suppose Z_k and S_j represent the time taken by k arrivals and by j phases in a given service interval respectively. Evidently, the conditional probability distributions of Z_k and S_i are of the Erlang type. Without loss of generality, all service times and the arrival process are assumed to be statistically independent. The following notations will be used throughout the paper. Additional notations will be introduced when necessary.

Notations

$G(t)$	: a finite mixture of Erlang distribution.
$F(t)$	: a general distribution function.
$F(dy)$	: a density function, i.e. $d\Pr\{Y \le y \mid m = (i, j, k)\}$.
$p(i, j, k)$	: a steady state probability of state (i, j, k).
X, X_i	: the random variable drawn from the stream of $G(\cdot)$.
Y	: a random variable drawn from the stream of $F(\cdot)$.
$\Gamma_{l,g}(t)$	: an Erlang distribution with parameters l and g,

$$\Gamma_{l,g}(t) = 1 - \sum_{h=0}^{g-1} e^{-lt} \cdot \frac{(lt)^h}{h!}.$$

S_j	: a random variable of $\Gamma_{l',j}$, i.e., time spent in completing j phases.
Z_k	: a random variable of $\Gamma_{l,k}$, for $1 \le k \le N-1$.
$\Pr\{r, r'\}$	: a transition probability from state r to state r'.

Let t_n, $n = 0,1,2,\ldots$, be the times for the customers departing from server 2. Consider the process (i_n, j_n, k_n) that embedded at t_n, $n = 0,1,2,\ldots$, where $i_n\,(k_n)$ is the number of the customers presently at station $1\,(2)$, and j_n is the number of phases left in the current service time of server 1 where $j_n = 0$ denotes server 1 is idle. Clearly, the domain of (i_n, j_n, k_n) is that of $0 \le i_n \le N-1$, $0 \le j_n \le M$, and $0 \le k_n \le N-1$. Note that by definition, it has $0 \le i_n + k_n \le N-1$ and $i_n = 0$ implies $j_n = 0$. Let R be the set of states of this system. Then the process is modeled as an embedded Markov chain with state space R because the Poisson arrivals and the linear combination of exponential distributions of service times are assumed. The state space R of this process is composed of the union of the following four classes:

i. $C_1 = \{(0,0,0)\}$: i.e., there are no customer in each station.

ii. $C_2 = \{(0,0,k_n) : 1 \le k_n \le N-1\}$: i.e., there are only k_n customers in station 2, and the staion 1 is empty.

iii. $C_3 = \{(i_n, j_n, 0) : 1 \le i_n \le N-1; 1 \le j_n \le M\}$: i.e., there are i_n customers in station 1, and j_n phases left in the current servic time in station 1.

iv. $C_4 = \{(i_n, j_n, k_n) : 1 \le i_n,\ 1 \le k_n$ and $i_n + k_n = N\text{-}1,\ 1 \le j_n \le M\}$: i.e., there are i_n customers in station 1 and k_n customers in station 2, and j_n phases left in the current service time of server 1.

Thus $R = C_1 \cup C_2 \cup C_3 \cup C_4$.

The total number of the elements in R for each $n > 0$, is calculated by

$$1 + (N-1) + M \cdot (N-1) + M \cdot H_{N-3}^3, \text{ where } H_{N-3}^3 = \frac{(N-1)(N-2)}{2}.$$

Because transitions in R is complicated, we will not write out the transition probabilities in matrix form explicitly. Instead, the transitions of states will be given and the conditions of joint probabilities will be discussed and presented in Section 3. Next, we should claim the embedded Markov chain at this model is irreducible and aperiodic.

Lemma 1. *The embedded Markov chain* $\{(i_n, j_n, k_n), \quad n = 1,2,3,\ldots\}$ *is irreducible and aperiodic.*

Proof.

(i) For each n >0, consider three cases according to their classes in R. The following diagrams (Figures 2-4) illustrate the irreducible property: First, in each figure, let $(i,j,k) \to (l,m,n)$ denote the transition from state (i,j,k) to state (l,m,n) with probability greater than 0, for every pair $(i,j,k), (l,m,n) \in R$. In Figure 2 we illustrate the transitions among four different classes. In Figure 3 (4) we illustrate the transitions between different states in C_3 and C_4 respectively. Therefore all pairs of the states are communicative, which implies the Markov chain is irreducible.

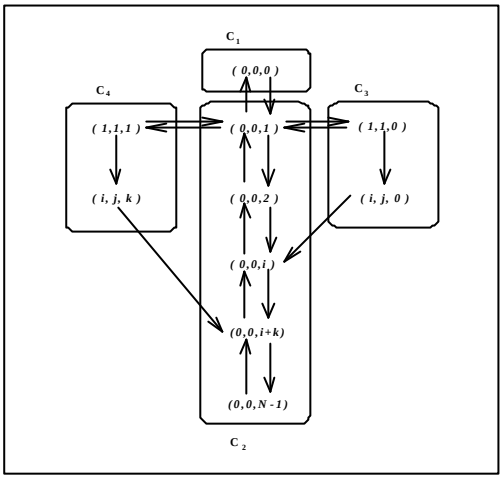

Figure 2.

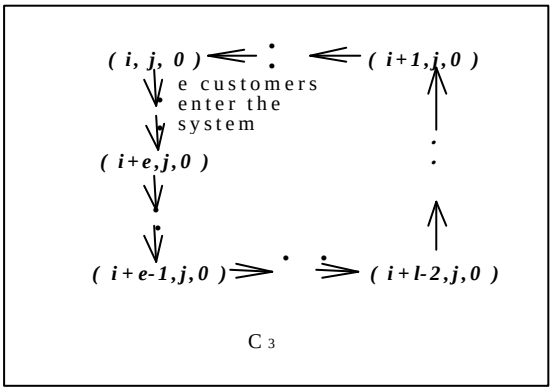

Figure 3.

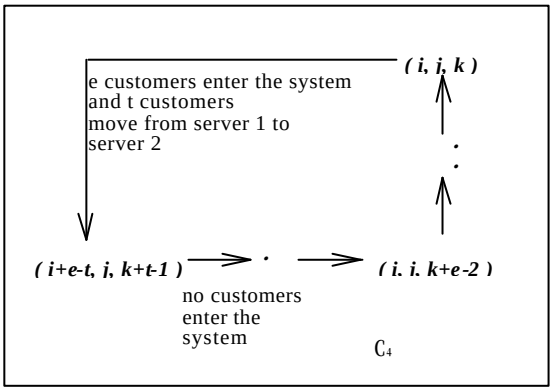

Figure 4.

(ii) The probability that each state in R returns to itself is greater than 0 for $F(0^+)<1$ and $G(0^+)<1$. The greatest common divisor of periods that each state returns to itself is 1. Thus this chain is aperiodic. As a result, the Markov chain is irreducible and aperiodic. Because the number of states in the embedded Markov chain in this system is finite, all states are positive recurrent.

3. Equilibrium Distributions of the Embedded Markov Chain

Since the system is finite, the steady state is achievable as long as the Markov chain is aperiodic. Let

$$P = \left[\Pr\{r, r'\} \right], r, r' \in R.$$

be the transition probability matrix of the embedded Markov chain. We shall determine P, and claim the Markov chain is ergodic. For achieving this aim, we define some independent random variables and corresponding distributions in the following.

Define：

$$s_j = \Pr\{S_j < Y\}, \tag{1}$$

$$g_j = \Pr\{Z_j < Y\}, \tag{2}$$

$$a_i = \Pr\{Z_i < X\}, \tag{3}$$

$$b_{i,j} = \Pr\{Z_i < S_j\}, \tag{4}$$

$$u_{m,n} = \Pr\{\sum_{i=1}^{m} X_i + S_n < Y\}, \tag{5}$$

$$\bar{s} = \Pr\{S_1 > Y\}, \tag{6}$$

$$\bar{g} = \Pr\{Z_1 > Y\}, \tag{7}$$

$$\bar{b_j} = \Pr\{Z_1 > S_j\}, \tag{8}$$

$$\bar{a} = \Pr\{Z_1 > X\}. \tag{9}$$

Here we introduce derivation of transition probabilities as follows.

Lemma 2. *By Equations* (1)-(9) *and assumptions for steady states, we have the following transition probabilities:*

(a) $\Pr\{(0,0,0),(i',j',k')\}$

$$= \sum_{i=1}^{N-2} \sum_{h=1}^{M} \Pr\{(i,h,1),(i',j',k')\} q_h (a_i - a_{i+1}) + \Pr\{(0,0,1),(i',j',k')\} \bar{a}.$$

(b) $\Pr\{(i,j,0),(i',j',k')\}$ $\qquad 1 \le i \le N-1; 1 \le j \le M$

$$= \begin{cases} \sum_{m=1}^{N-i-1} \sum_{h=1}^{M} \Pr\{(m+i-1,h,1),(i',j',k')\} q_h (b_{m,j} - b_{m+1,j}) \\ \quad + \sum_{h=1}^{M} \Pr\{(i-1,h,1),(i',j',k')\} q_h \bar{b_j} \end{cases} \quad \text{if} \quad i \ge 2,$$

$$\begin{cases} \sum_{m=1}^{N-2} \sum_{h=1}^{M} \Pr\{(m,h,1),(i',j',k')\} q_h (b_{m,j} - b_{m+1,j}) \\ \quad + \Pr\{(0,0,1),(i',j',k')\} \bar{b_j} \end{cases} \quad \text{if} \quad i = 1.$$

(c) $\Pr\{(0,0,k),(i',j',k-1)\}$

$$= \begin{cases} \bar{g} & \text{if} \quad i' = j' = 0, \\ (g_{i'} - g_{i'+1}) \{ \sum_{h=j'+1}^{M} (s_{h-j'} - s_{h-j'+1}) q_h + \bar{s} \cdot q_{j'} \} & \text{if} \quad N-k > i' > 0, \\ g_{N-k} \{ \sum_{h=j'+1}^{M} (s_{h-j'} - s_{h-j'+1}) q_h + \bar{s} \cdot q_{j'} \} & \text{if} \quad i' = N-k. \end{cases}$$

(d) $\Pr\{(i,j,k),(i',j',k)\} \qquad i' \geq i, \quad i' \neq 0$

$$= \begin{cases} (g_{i'-i+1} - g_{i'-i+2}) \sum_{h=j'}^{M} (s_{h-j'+j} - s_{h-j'+j+1}) q_h & \text{if} \quad i' < N-k-1, \\ g_{N-k-i} \sum_{h=j'}^{M} (s_{h-j'+j} - s_{h-j'+j+1}) q_h & \text{if} \quad i' = N-k-1. \end{cases}$$

(e) $\Pr\{(i,j,k),(0,0,k')\} \qquad k < k', \quad i \geq 1$

$$= \begin{cases} \bar{g}(u_{i-1,j} - u_{i-1,j+1}) & \text{if } i+k = k'+1, \\ (g_{k'-i-k+1} - g_{k'-i-k+2})(u_{k'-k,j} - u_{k'-k,j+1}) & \text{if } i+k < k'+1, \\ g_{N-i-k}(u_{k'-k,j} - u_{k'-k,j+1}) & \text{if } k' = N-1. \end{cases}$$

(f) $\Pr\{(i,j,k),(i',j,k\text{-}1)\} \qquad i' \geq i$

$$= \begin{cases} \bar{s} \cdot \bar{g} & \text{if} \quad i' = i, \\ \bar{s}(g_{i'-i} - g_{i'-i+1}) & \text{if} \quad N-k > i' > i, \\ \bar{s} \cdot g_{N-k-i} & \text{if} \quad i' = N-k. \end{cases}$$

(g) $\Pr\{(i,j,k),(i',j',k')\} \qquad 1 \leq k < k', \quad i' \geq i$

$$= \begin{cases} (g_{k'-k+i'-i+1} - g_{k'-k+i'-i+2}) \left\{ \sum_{h=j'+1}^{M} (u_{k'-k,h-j'} - u_{k'-k,h-j'+1}) q_h + \bar{s}\, q_{j'} \right\}, \\ \qquad \text{if } i' \neq N-k'-1, \\ (g_{N-k-i}) \left\{ \sum_{h=j'+1}^{M} (u_{k'-k,h-j'} - u_{k'-k,h-j'+1}) q_h + \bar{s}\, q_{j'} \right\}, \text{ if } i' = N-k'-1. \end{cases}$$

(h) $\Pr\{(i,j,k),(i,j',k\text{-}1)\} \qquad 1 \leq k; j' < j$

$= (s_{j-j'} - s_{j-j'+1})\bar{g}$.

(i) $\Pr\{(i,j,k),(i\text{-}1,j',k)\} \qquad i \geq 1$

$$= \begin{cases} \bar{g} \sum_{h=j}^{M} (s_{h-j'+j} - s_{h-j'+j+1}) q_h & \text{if } i \geq 2, \\ \bar{g}\,(s_j - s_{j+1}) & \text{if } i = 1, \ j' = 0. \end{cases}$$

Proof.

(a) Given a state $(1, j, 0)$, it is with probability 1 that the transition from state (0, 0, 0) to state $(1, j, 0)$ for some j. Note that server 2 remains idle until an arrival occurs and moves to station 2 from station 1. If there are no customers occur during his service time at station 1, the probability of reaching state (i', j', k') is given by $\Pr\{(0, 0, 1),(i', j', k')\}$. Otherwise, it leads to $\Pr\{(i, h, 1),(i', j', k')\}$ when there are exactly i arrivals during his service at station 1 and the service time at station 1 is Erlang-h.

(b) There are two cases to be considered as discussed in (a). If there are no arrivals occurring during the service of Erlang-j, the probability of reaching state (i', j', k') is given by $\Pr\{(0, 0, 1),(i', j', k')\}$ when $i = 1$, otherwise it is $\Pr\{(i - 1,h,1),(i', j', k')\}$ when $i \geq 2$. The similar arguments are applied to the case of more arrivals occurring during the service time of Erlang-j.

(c) If $i' = j' = 0$, then no arrival occurs during the service time at station 2. The probability of it is $\bar{g}$. If $i' > 0$, then during the service time Y there have arrived i' customers and one of i' customers moved $h - j' + 1$ phases in service given that it requires a service time of Erlang-h. If it requires a service time of Erlang-j' for $j' > 0$, then with probability $\bar{s}\, q_{j'}$ it remains in phase j' when Y is completed. Given the total number of tokens available for arrivals, multiplication of the probabilities of number of arrivals and number of phases in summing up for all possible h gives the results required. This is because of assumption of independence of arrivals and the service times.

(d) If $i' < N - k - 1$, then there have arrived $i' - i + 1$ customers during the service time Y and no lost customers. Otherwise, the number of arrivals may exceed the number of tokens available. Given h in service time, the total of number of phases in service is j phases plus h-j phases provided that a current service time is of Erlang-h. Applying the independence assumption of service times and arrivals again, the probabilities are calculated based on the condition of a number of arrivals.

(e) If $i + k = k' + 1$, then no arrival occurs during the service time Y and i customers complete their services at station 1. It includes that the first one of i customers finishes j phases and exactly i-1 customers have completed entire services at station 1. The probability of it is $\bar{g}(u_{i-1,j} - u_{i-1,j+1})$. If $i - k < k' + 1$, then there are $k' - i - k + 1$ arrivals and there are $k' - k + 1$ customers in total move to station 2 during the service time at station 2. Notice that exactly $k' - k$ of them have completed entire services. If $k' = N - 1$, the similar arguments applies.

(f) There are no arrival and no finishing in service at phase j during the service time Y in case of $i' = i$. If $i' > i$, there are $i' - i$ arrivals but it does not finish service at phase j during Y. If $i' = N - k$, the similar arguments applies.

(g) There are $k' - k + i' - i + 1$ customers entering the system. Only $k' - k - i + 1$ of them have reached station 2 from station 1. Thus the same arguments in (e) are applied. However, it must take into account all possible service phases h. In the case of $i' \neq N - k' - 1$, there are $h - j' + 1$ phases of service during Y as it has the service time of Erlang-h at station 1. The probabilities are computed according to the number of tokens available, i.e., the effective arrivals.

(h) There are no arrivals during the service time Y and it has moved exactly $j - j' + 1$ phases during it.

(i) There are no arrivals and exactly one customer moves to station 2 from station 1 during Y. If i =1, then the customer moving from station 1 to station 2 has finished service of j phases; otherwise, after he finishes his service the next customer will move to phase j' given its service time is Erlang-h.

Lemma 2 defines the transition probabilities within R. From the definition of states in the model, the steady-state probabilities p exist. The steady-state balance equations for P are given by

$$p(r)P = p(r) \text{ and } \sum_{r\in R} p(r) = 1 \text{ where } r \in R. \tag{10}$$

Let $p(r)$ be the steady-state probability of state r in the system where $r = (i, j, k)$. Unfortunately it does not posses a formula-type solution of $p(i, j, k)$. Based on Equations (1)-(9), it involves multiple integrations in obtaining transition probabilities of entries in P. For solving it, numerical solving techniques should be employed with present computing facilities. In addition, it would be also desirable to make use of the fact that coefficient matrix has a large number of zero entries where a sparse matrix solution procedure is appropriate. However, all of these associated with particular numerical solving skills are not our goal in this paper and thus are not discussed in detail here.

In general, π may be considered as a function of N. Given the size of N, in terms of a number of tokens available, Procedure 1 in the following determines the joint probabilities.

Procedure 1. A solution procedure for p

Step 1: Compute the Probabilities (1)-(9).

Step 2: Attain p by solving (10).

In short, after p is solved with a standard solving Procedure 1, several performance measures may be derived by using p. For example, the probability of idle of server 2 is

$$\sum_{i=1}^{N-1}\sum_{j=1}^{M} p(i,j,0) + p(0,0,0).$$

Similarly, a probability of idle time of server 1 is

$$\sum_{k=1}^{N-1} p(0,0,k) + p(0,0,0).$$

Other conventional measurements are easy to obtain by basic queueing formulas.

However, the loss probability of customers are different from existing formulas since the system has general service time. In the sequel, we shall describe how to derive the loss probabilities by p.

4. Probability of Loss of Arrivals

The definition of the loss probability in our model is the probability of that an arrival who finds the token buffer empty is rejected by the system. From Section 3 we derive the joint distribution considered at the departure points. By the definition, state (i, j, k) means there are $(N - i - k)$ tokens in the token buffer that the departure customer observes including the one which was just released. It is clear $N - i - k \geq 1$. Now let h_n, $n \geq 1$ represent the probability that a departure finds there are n tokens in the token buffer. Thus we have the following formulas of h_n, $n \geq 1$:

$$h_n = \sum_{\substack{i+k=N-n \\ 1\le i,k\le N-n-1 \\ 1\le j\le M}} m(i,j,k) + \sum_{1\le j\le M} m(N-n,j,0) + m(0,0,N-n),\ n\ge 1.$$

In particular, when $n=1$ we have

$$h_1 = \sum_{\substack{i+k=N-1 \\ 1\le i,k\le N-2 \\ 1\le j\le M}} m(i,j,k) + \sum_{1\le j\le M} m(N-1,j,0) + m(0,0,N-1),$$

which is the probability of a departing customer observing one token in the token buffer. Notice that what we are interested in is the probability an arrival finds no token, i.e., a probability of losing customers. Obviously, h_1 is not the loss probability. However, this is the hint for us to proceed.

Notice that when an arrivals finds no token in the token buffer the system must have passed through the state where there was only one token in the token buffer. The other fact is that there is at least one token seen by a departure whenever it leaves the system. Then we divide all states into two parts to study the loss probability:

(i) Consider state (i,j,k), $k\ge 1$, i.e., there is at least one customer in the station 2. If there are more than $N-i-k+1$ customers entering the system before a customer in station 2 completes his service, the probability of losing one or more arrivals occurs.

(ii) Consider state $(i,j,0)$, i.e., there are no customers in station 2. If there are more than $N-i$ customers arrive at the system during the service times of j phases, then we have a probability of losing one or more arrivals. Based on the arguments above, a theorem of the loss probability is stated as follows:

Theorem. *The loss probability* q *is derived by*

$$q = \sum_{m=2}^{N-1}\sum_{k=1}^{m-1}\sum_{j=1}^{M} p(m-k,j,k)\cdot g_{N-m+1} + \sum_{k=1}^{N-1} p(0,0,k)\cdot g_{N-k+1}$$

$$+ \sum_{i=1}^{N-1}\sum_{j=1}^{M} p(i,j,0)\cdot \Pr\{Z_{N-i+1} < Y + S_j\} + p(0,0,0)\cdot \Pr\{Z_{N+1} < Y + X\}.$$

Proof. The first term of right hand side is written for the case when $k\ge 1$ and $i\ge 1$. In this case the loss probability occurs when there are more than $N-i-k$ customers come to the system. The probability of more than $N-i-k$ customers come to the system is $\Pr\{Z_{N-i-k+1} < Y\}$, namely $g_{N-i-k+1} = g_{N-m+1}$ as $m = i+k$. This probability is considered for all $i+k\le N-1$. When $i=0$, implying $j=0$, it leads to the second term. If $k=0$, the service time at station 1 should be taken into account. Depending on the service times at stations 1 and 2, a lost customer occurs when a number of arrivals during the service time exceeds the number of tokens available in the token buffer. According to the initial state, we have the third and the fourth term.

Notice that the loss probability is a function of N. From the derivation of it in Theorem, it decreases as N increases. The existence of such an N is immediately clear since the domain of N defined in our paper is not empty and finite.

The above probability distribution is a function of the decision variable N. The distribution exhibits certain monotonicity properties in relation to N. This will play a

crucial role in the optimization part of this study. We can now state the basic property of monotonicity for the steady state behavior of the system.

Corollary 1. Let $L_i(N)$ be the expected number of customers at station i provided there are N tokens in the system, where $N < \infty$. Then, we have

(a) $L_1(N) \le L_1(N+1)$,

(b) $L_2(N) \le L_2(N+1)$.

Proof. Because the service times at stations 1 and 2 are independent and nonnegative, we discuss (a) only. Since the arrival process is Poisson, the number of arrivals grows stochastically nondecreasingly as the system flow time increases; so does the number of customers in system. Hence, the expected number of customers in system follows.

5. An Optimization Model

We are interested in minimizing a cost function, $C(N)$, in term of an expectation of a number of customers at each station plus the maintenance cost of N tokens subject to the constraint that the loss probability must not exceed a certain number and the throughput must cross over above a specified level. Denote by $q(N)$ and $T(N)$ with respect to the loss probability and the mean throughput with N tokens in the system. Then it is correspondingly equivalent to develop a procedure of minimizing $C(N)$ subject to $q(N) < w_1$, $T(N) > w_2$, where w_1 is the tolerance of loss probability and w_2 is the minimal level of throughput. Before considering optimization, we verify the monotonicity of $q(N)$ and $T(N)$.

Corollary 2. Under the steady-state assumptions, we have

(1) $q(N)$ is monotonically nonincreasing of N, i.e., $q(N) \ge q(N+1)$.

(2) $T(N)$ is monotonically nondecreasing of N, i.e., $T(N) \le T(N+1)$.

Since the expected number of customers during a specified time interval will converge, the corollary is clear by using the fact of Theorem and Corollary 1. It shows that the number of throughput increases in average when the total number of tokens is increased by one. This is easily verified to be true by modeling the system with a closed cyclic model of tokens and studying its cycle time. The distribution of the total number of tokens at each station will be identical to that of a closed cyclic system with N customers. The work that follows is an extension of Luh's model [3]. The proof is omitted here.

In particular, the mean throughput of this system can be written as $l(1 - q(N))$. If $w_1 \le 1 - w_2 / l$, let $w = w_1$; otherwise, let $w = 1 - w_2 / l$. To determine N^*, we begin with $N = 1$. If $q(N) > w$, then increasing N by 1 will reduce the value of q. Checking the value of $q(N+1)$, it reiterates the procedure until $q(N+1) \le w$. Let $N^* = N + 1$. We propose a heuristic method which minimizes $C(N)$ by taking Theorem into account. Moreover, because N^* minimizing $C(N)$ may not be an integer, Procedure 2 that searches for an integer $\bar{N}$ which is suboptimal to N^* is described as below.

Procedure 2. Search for $\bar{N}$:

Step 0: Let $N = 1$.

Step 1: Compute $q(N)$.

Step 2: If $q(N) \le w$, then go to step 3; else let $N = N + 1$ and go to step 1.

Step 3: Print $\bar{N} = N$ and stop.

Given the system data, 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 reaches a suboptimal point. Since in practice a domain of continuous values of N is not permitted, the resulting value serves as the least upper bound among integers 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 is involving evaluating the probability distributions of the system states. The difficulty just indicated and the fact that integer values are required for the decision variables, suggest the use of an implicit enumeration scheme for the optimization algorithm.

An Example

The following problem was run on a PC using Mathematica 4.0 to calculate Equations (1)-(9) needed to yield the joint stationary probabilities. Consider the following optimization problem:

Minimize $c_1 L_1 + c_2 L_2$
Subject to $q(N) \le w_1$, $T(N) \ge w_2$, $N > 0$, Integer.

The optimization aspect of this study is a formidable one because an expression for the availability as a function of the decision variable does not exist in closed algebraic form. That is, $q(N)$ that appears in the constraint is determined from Theorem and can only be calculated numerically when the value of N is specified. The parameters were set as follows: $c_1 = 1$, $c_2 = 2$, $l = 18$, $w_1 = 0.1$, $w_2 = 17$,

$$F(t) = \int_0^t \frac{1}{0.02x\sqrt{2\mathrm{p}}} e^{\frac{-1}{2}\left(\frac{\ln x - 0.05}{0.02}\right)^2} dx, \quad t \ge 0,$$

$$G(t) = 1 - \sum_{h=0}^{4} \frac{(t/4)^h e^{-t/4}}{h!}, \quad t \ge 0.$$

Consider two different service distributions with common mean service time 0.05 at each station. The probability functions at stations 1 and 2 are of the Erlang type with 5 phases and lognormal with standard deviation 0.02, respectively. By applying procedure 1 and 2, we obtain $\bar{N} = 10$.

For an illustrative purpose, the service time $G(t)$ at station 1 is fixed for all test problems. We present numerical experiments describing the various aspects of system behavior in different values of arrival rate l. Alternatively, the service times at station 2 are assumed to be of the Erlang type with 2 phases and lognormal distributions for comparison. According to three arrival rates, l =18, 22 and 27, the probability of loss of arrivals is computed in each example, tested with different token capacities from 5 to 35. This is demonstrated in Figure 5.

In Figure 5, it shows the nonincreasing property of the probability of loss of arrivals which is independent of the statistical distributions of the interarrival and service time distributions. The solid line represents the case of lognormal distributions while the dot lines depict that of the Erlang distributions. Observe that in terms of the loss probability, lognormal distributions outperform that of all Erlang distributions statistically. But the probability of loss shows steeply dropping before N = 10 and moving down slowly and smoothly afterwards regardless of the type of service time distributions. It may be

explained in the following. The expected total service time at the network is at least 0.1 units of time. Equivalently, there are expectedly 10 customers per unit time completing services at the network. When the number of tokens is less than 10, the probability of no token at token buffer is comparatively high. Conversely, when it exceeds 10, the probability of loss approaches to its limit closely, but it converges slowly in distribution. Now, consider the effect of the traffic load. Apparently, the heavier the traffic, the higher the probability of loss becomes as shown in Figure 5. Notice the average cost of $l = 22$ and $l = 27$ grows almost indifferently since neither of them has reached the limit as shown in Figure 6.

In Figure 6, we illustrate the average cost behavior which grows nondecreasingly versus the number of tokens. Since the probability of loss approaches to zero as N is greater than 10 when $l = 18$, the expected number of customers at network converges to a constant; so does the average cost.

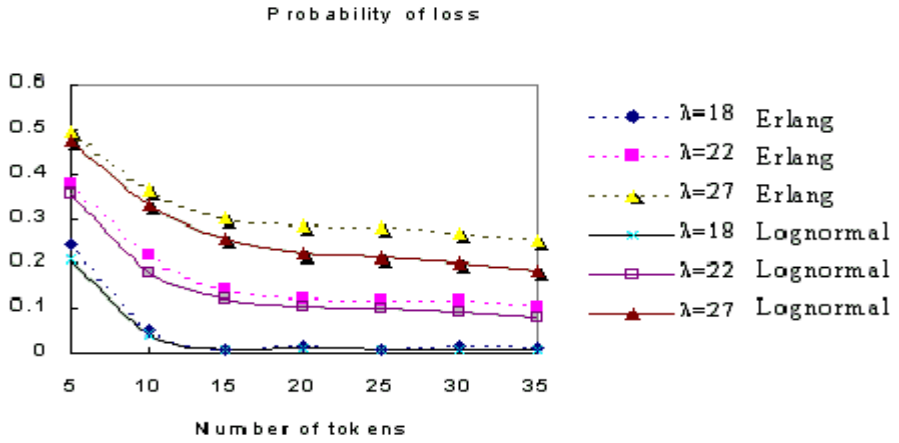

Figure 5. Erlang_2 and Lognormal service time.

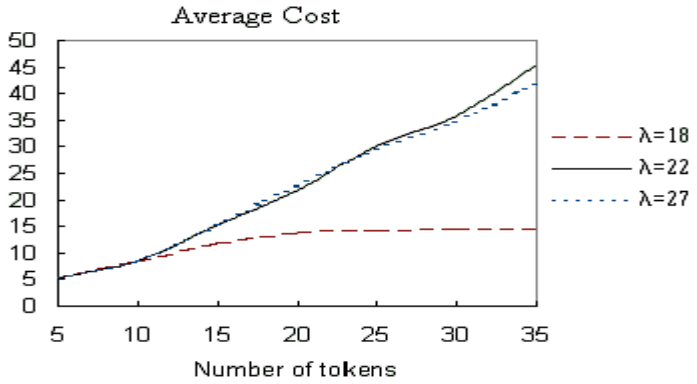

Figure 6. Cost versus number of tokens.

6. Conclusions

We have extended the closed queueing network in Luh [3] to an open queueing system with a token buffer, providing a loss probability. The expressions of our recursion scheme can be differentiated readily to obtain all existing moments. Furthermore, the expected cost may be minimized subject to constraints on the loss probability and throughput. This modeling and analysis has provided a solution to other semi-Markovian representation of performance measures of general servers in queueing networks.

We have shown in the stochastic process analysis that the monotonicity conditions hold. It turns out the probability of loss has a monotonic property. Note that although the optimization technique is illustrated with a linear cost function, it appears without modification to any cost function that is monotone in the decision variable since (L_1, L_2) are monotonic functions of N. That suggests the system manager how to provide a method to solve a system with tokens that is complicated and cost effective.

References

1. Buzacott, J. A. and Shanthikumar, J. G. (1993). *Stochastic Manufacturing Systems*, Chapter 5. Prentice-Hall International Edition.
2. Daduna, H. (1985). Two-stage cyclic queues with nonexponential severs: steady-state and cyclic time. *Operation Research*, 34(3), 455-459.
3. Luh, H. (1999). Derivation of the N-step interdeparture time distribution in $GI/G/1$ queueing systems. *European Journal of Operational Research*, 118, 194-212.
4. Miller, L. W. (1975). A note on the busy period of an $M/G/1$ finite queue. *Operations Research*, 23, 1179-1182.

Authors' Biographies:

Hsing Luh is a Professor at Department of Mathematical Sciences, National Chengchi University, Taiwan. His research interests include queueing theory and applications, stochastic models and simulations, as well as linear programming and optimization.

Chun-Lian Huang received his MS degree from Department of Mathematical Sciences, National Chengchi University. Currently, he is a high school teacher.