Analysis of an Open Tandem Queueing Network with Population Constraint and Constant Service Times

Y. Rhee
H. G. Perros

Center for Communications and Signal Processing
Department of Computer Science
North Carolina State University

TR-94/11
May 1994

Analysis of an Open Tandem Queueing Network with Population Constraint and Constant Service Times [1]

Young Rhee

Graduate Program in Operations Research and
Center for Communications and Signal Processing
North Carolina State University
Raleigh, NC 27695-7913

Harry G. Perros

Computer Science Department and
Center for Communications and Signal Processing
North Carolina State University
Raleigh, NC 27695-8206

[1] supported in part by a grant from IBM-RTP

ABSTRACT. We consider an open tandem queueing network with population constraint and constant service times. The total number of customers that may be presented in the network can not exceed a given value K. Customers arriving at the queueing network when there are more than K customers are forced to wait in an external queue. The arrival process to the queueing network is assumed to be arbitrary.

We show that this queueing network can be transformed into a simple network involving only two nodes. Using this simple network, we obtain an upper and lower bound on the mean waiting time. These bounds can be easily calculated. Validations against simulation data establish the tightness of these.

KEY WORDS: open queueing network, population constraint, constant service times, semaphore queue, upper and lower bounds.

1. Introduction

In this paper, we consider an open tandem queueing network with constant service times and population constraint. The total number of customers simultaneously present in the network can not exceed a given value K. Customers arriving while the network has K or more customers are forced to wait in an external queue which is assumed to have an infinite capacity. As soon as a customer leaves the network, the first customer in the external queue is allowed to enter the network. Queueing networks with population constraints do not have a closed form solution. As a result, several approximate solutions have been proposed in the literature.

In a population constrained queueing network, the number of customers in the network can be controlled by a semaphore queue. A semaphore queue consists of an external queue for customers waiting to enter the queueing network and a token pool for unused tokens. A customer can not enter the queueing network unless it has a token. A customer arriving at the queueing network to find the token queue empty, is queued in the external queue. Upon departure of a customer from the queueing network, the token is returned back to the pool. At that moment, if there is at least one customer waiting in the external queue, the first customer takes the token and enters the queueing network. The semaphore queue mechanism was first used to model window flow control mechanisms in Reiser [19]. Fdida, Perros and Wilk [4] presented an approximation technique for

solving an open queueing network with Poisson arrivals and exponential service times which consists of several subnetworks where each subnetwork is controlled by semaphore queue. They used this type of network to model the nested sliding window flow control mechanisms. Dallery [3] analyzed a single class open queueing network with population constraint assuming Coxian service times. He first transformed the basic open model into an equivalent closed model by exchanging the roles of customers and tokens. The model obtained was a closed queueing network whose population is equal to the maximum allowable number of customers in the open queueing network. The closed queueing network was then analyzed using Marie's approximation method [14]. Shapiro and Perros [21] presented a hierarchical method for analyzing nested sliding window flow control mechanisms with packet fragmentation and reassembly. An approximation algorithm was presented to hierarchically reduce the network to a single queue whose performance characteristics represented the original network. Perros, Dallery and Pujolle [18] extended the approach proposed in Dallery [3] in order to analyze open multiclass queueing networks with class dependent population constraints. Other semaphore controlled queueing models have been also considered in the literature. For two node queueing networks with population constraint see Perros [17] and references within. Lam [10] extended the class of multichain queueing networks of the product-form type to include mechanisms of state dependent lost and triggered arrivals. Kaufman and Wang [7] analyzed a queueing network with Poisson

arrivals and exponential service times. They derived the stability condition and proposed an analytic approximation for the mean waiting time.

Several studies of open tandem queueing network with constant service time and without a population constraint have been reported in the literature. Avi-Itzhak [2] and Friedman [5] analyzed an open tandem configuration with blocking and constant service times, assuming that the first queue has an unlimited capacity. Altiok and Kao [1] presented a lower and upper bound on the throughput of the same queueing network assuming that the first queue is finite. Ziegler and Schilling [22] and Gall [6] studied the delay decomposition in a queueing network with Poisson arrivals and constant service times and assume to be infinite capacity queue. Shalmon and Kaplan [20] considered a tandem network with constant service times and multiple interfering sources. Assuming that all nodes have an infinite capacity, they derived the steady-state moment generating function for the waiting time. They also presented a complete delay analysis for an embedded $M/G/1$ queue and a batch arrival $M/G/1$ queue. Finally, Newell [16] analyzed a tandem network with constant service times, assuming that the first node is saturated.

The paper is organized as follow. In section 2, we persent an open queueing network with constant service times and population contraint. We also obtain a two node queueing network which is equivalent to the original queueing network as far as a customer's waiting time is considered. In section 3, we give an upper

and lower bound of the mean waiting time in the queueing network. An improved upper and lower bound is also described. In section 4, we establish the tightness of these bounds using simulation results. Finally, the conclusions are given in section 5.

2. An open tandem queueing network with population constraint.

Let us consider an open tandem queueing network with a population constraint and constant service times. We assume that the queueing network consists of N nodes. The arrival process to the queueing network is assumed to be a generally independent with rate λ and the service time at each node is constant equal to $s_i, i = 1, 2 \ldots N$. The population constraint of the queueing network is controlled by a semaphore as shown in Figure 1.

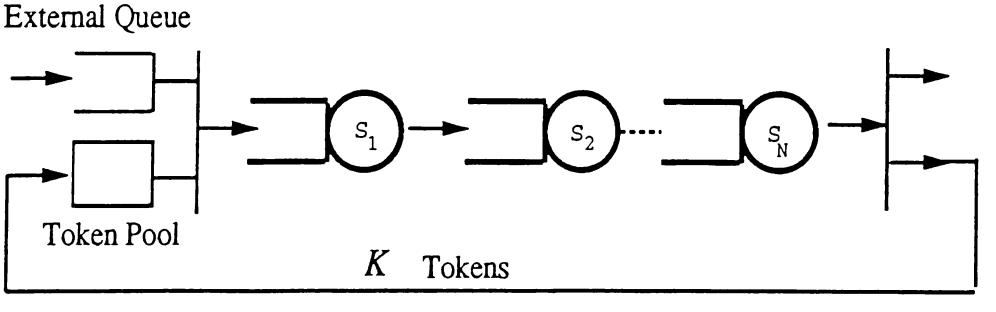

Figure 1: An open queueing network with constant service times

The semaphore mechanism consists of a pool of K tokens and an external queue. An arriving customer takes a token and enters the queueing network. The customer holds this token until it leaves the network. At that time, the token is

returned to the pool in zero time. Customers that arrive during the time when the pool is empty are forced to wait in the external queue. The first customer in the external queue enters the queueing network as soon as a token is returned to the pool.

For this queueing network, the waiting time of a customer remains the same even though the order of the service times is rearranged.

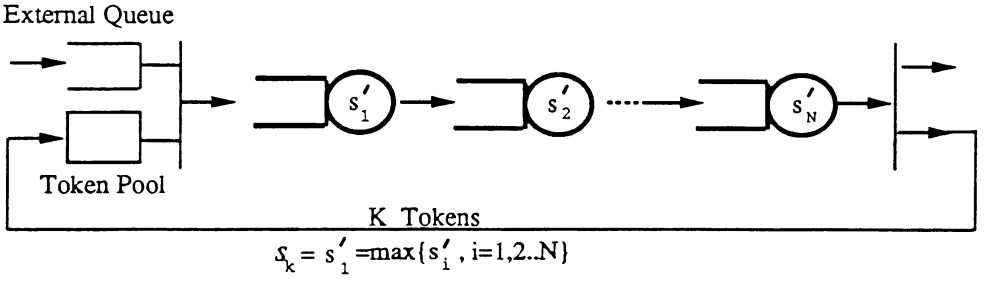

Figure 2: Rearranged open queueing network with constant service times

Theorem 1. Let us rearrange the original open queueing network into an open queueing network where the node with the longest service time is placed at the begining of the queueing network as shown in Figure 2. A customer's waiting time in either queueing network is the same.

Proof . Let w_{pj} be customer p's waiting time at node j, $j = 1, 2, \ldots N$. Also, let a_{pj} be the interarrival time between the p^{th} and the $(p-1)^{st}$ customer to node j, $j = 1, 2 \ldots N$, and a_{p0} be customer p's interarrival time to the external queue. Finally, let $s_k = max\{s_j, j = 1, 2, \ldots N\}$, that is the k^{th} node is assumed to be the node with the longest service time.

We first consider the case that an arriving customer p in the original queueing network finds the token pool non-empty. We assume that arriving customers are numbered sequentially, and that p is the sequential number of our tagged customer. Let w_i^e be the i^{th} customer's waiting time in the external queue. Then customer p's interarrival time to the first node is $a_{p1} = a_{p0} - w_{p-1}^e$. If the first node is idle, then customer p does not wait at the first node, and the interarrival time to the second node is $a_{p2} = a_{p1} - w_{(p-1)1}$. However, if the first node is busy, then the interarrival time to the second node is $a_{p2} = s_1$ and customer p's waiting time at the first node is $w_{p1} = w_{(p-1)1} + s_1 - a_{p1}$. In general, for any node j ($j \leq k$), we have

$$a_{p(j+1)} = \begin{cases} s_j & \text{if } w_{pj} > 0 \\ a_{pj} - w_{(p-1)j} & \text{if } w_{pj} = 0 \end{cases} \quad (1)$$

$$w_{pj} = max\{0, w_{(p-1)j} + s_j - a_{pj}\} \text{ for } j \leq k \quad (2)$$

There is no waiting for customer p after node k, because the interarrival time is always larger than the service time of the subsequent nodes. The total amount of waiting for customer p, w_p, is

$$w_p = \sum_{i=1}^{k} w_{pi}$$

$$= \sum_{i=1}^{k} max\{0, w_{(p-1)i} + s_i - a_{pi}\} \quad (3)$$

which can be shown recursively to be equal to

$$w_p = max\{0, w_{(p-1)} + s_k - a_{p0}\} \tag{4}$$

We now observe that in the rearranged queueing network, customer p will only wait in the first node for a period of time equal to $max\{0, w_{(p-1)} + s_k - a_{p0}\}$.

Let us now consider the other case where an arriving customer p in the original queueing network finds the token pool empty. That is, there are K customers in the queueing network. Then, customer p has to wait in the external queue until the previous $(p-K)^{th}$ customer leaves. Customer p's arrival time to the external queue is $\sum_{i=1}^{p} a_{i0}$ and the $(p-K)^{th}$ customer's departure time is $\sum_{i=1}^{p-K} a_{i0} + w_{p-K} + \sum_{i=1}^{N} s_i$. Thus,

$$w_p^e = w_{p-K} + \sum_{i=1}^{N} s_i - \sum_{i=p-K+1}^{p} a_{i0} \tag{5}$$

Therefore, customer p's interarrival time to the first node is

$$a_{p1} = a_{p0} + w_p^e - w_{p-1}^e \tag{6}$$

Once customer p enters node 1, its waiting time and interarrival times can be calculated using expressions (1) and (2). Therefore, the same customer in the rearranged queueing network will wait for the same amount of time as in the original queueing network. □

The general expressions of the waiting time and interarrival time for the

customer p are as follows;

$$
\begin{aligned}
w_p^e &= max\{0, w_{p-K} + \sum_{i=1}^{N} s_i - \sum_{i=p-K+1}^{p} a_{i0}\} \\
a_{p1} &= a_{p0} + w_p^e - w_{p-1}^e \\
w_{pj} &= max\{0, w_{(p-1)j} + s_j - a_{pj}\} \text{ for } j \leq k \\
a_{p(j+1)} &= \begin{cases} s_j & \text{if } w_{pj} > 0 \\ a_{pj} - w_{(p-1)j} & \text{if } w_{pj} = 0 \text{ and } j \geq 1 \end{cases}
\end{aligned} \tag{7}
$$

Let us consider the rearranged queueing network shown in Figure 2. Since there is no queue after the first node, the time a customer spends in the remaining nodes is the sum of the service times $\sum_{i=2}^{N} s_i$. In view of this, we can represent the queueing network in Figure 2 by a simpler two-node queueing network as shown in Figure 3. For presentation purposes we shall refer to these two node as the first node and the second node. s^* represents the longest service time in the network and $\overline{s}$ is the sum of the remaining service times, i.e., $\overline{s} = \sum_{i=1}^{N} s_i - s^*$. The number of parallel servers at the second node is infinite. A customer's waiting time in the two-node queueing network is the same as in the rearranged queueing network, and consequently it is the same as in the original queueing network under study. Below, we shall concentrate on the two-node queueing network. For all customers, the interdeparture times from the first node are greater than or equal to s^*. This means that the customers at the second node have different residual service times and the difference between any two consequtive residual service times is greater than or equal to s^*. This observation gives rise to the following theorem,

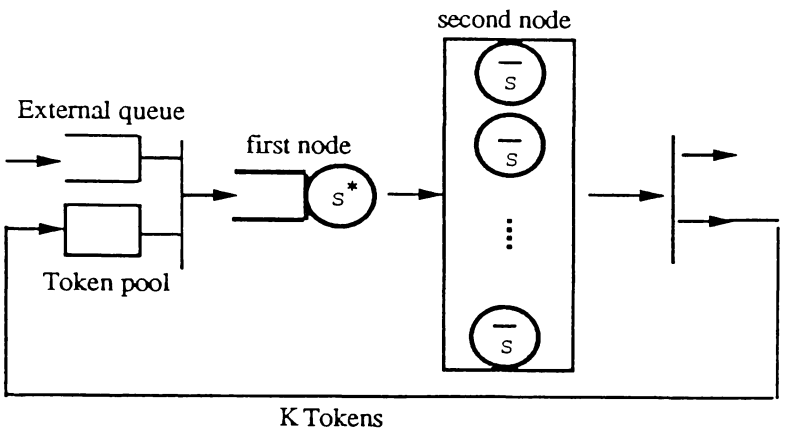

Figure 3: Two-node open queueing network with constant service times

Theorem 2. Let us consider the two-node queueing network as shown in Figure 3 and let K be the allowable number of tokens in the token pool and s^* be the service time at the first node and $\bar{s}$ be the service time at the second node. If $Ks^* \geq \sum_{i=1}^{N} s_i$, then a customer's waiting time in the queueing network is that of a $GI/D/1$ queue with a service time s^*.

Proof. Let us consider the instance $t^{(n)}$ when a new busy period of the queueing network begins. Let w_i , a_i and T be, the i^{th} customer's waiting time in the network, its interarrival time, i.e., the difference between the arrival times of the $(i-1)^{st}$ and i^{th} customer and the sum of all service times, i.e., $\sum_{i=1}^{N} s_i$ respectively. w_i can be divided into w_i^e and w_i^f , the i^{th} customer's waiting time at the external queue and the waiting time at the first node respectively. Customers arriving during a busy period are numbered sequentially starting from 1. Since $t^{(n)}$ is the

begining of a busy period, $w_1 = 0$, customer 1's departure time from the first node is $t^{(n)} + s^*$, and its departure time from the queueing network is $t^{(n)} + T$. There is no external waiting time for the first K^{th} customers, i.e. $w_i^e = 0, i = 1, 2 \ldots K$. For the j^{th} customer, $(2 \leq j \leq K)$, the waiting time in the first node, $w_j^f, j = 1, 2 \ldots K$, and its departure time from the first node, d_j^f, and its departure time from the queueing network, d_j, are given by the following expressions;

$$w_j^f = max\{0, w_{j-1} + s^* - a_j\} \tag{8}$$

$$d_j^f = t^{(n)} + \sum_{i=2}^{j} a_i + w_j + s^* \tag{9}$$

$$d_j = t^{(n)} + \sum_{i=2}^{j} a_i + w_j + T \tag{10}$$

The time when the first token returns to the pool is $t^{(n)} + T$, which is equal to d_1. The time when the first node is free for the $(K+1)^{st}$ customer is $t^{(n)} + \sum_{i=2}^{K} a_i + w_K + s^*$, i.e., d_K^f, which is also equal to $t^{(n)} + \sum_{i=1}^{K} w_i + Ks^*$. Finally, the time when the $(K+1)^{st}$ customer arrives to the external queue is $t^{(n)} + \sum_{i=1}^{K+1} a_i$. Since we assume $Ks^* \geq \sum_{i=1}^{N} s_i$, then $d_K^f \geq d_1$, which means the waiting time depends on the time when the first node is free for the $(K+1)^{st}$ customer only. In other words, once an arriving customer experiences the external queue waiting time, then the customer has to wait in the first node. For the $(K+1)^{st}$ arriving customer, the external queue and the first node waiting times are

$$w_{K+1}^e = max\{0, T - \sum_{i=2}^{K+1} a_i\}$$

$$w^f_{K+1} = max\{0, w_K + s^* - a_{K+1}\} - w^e_{K+1} \quad (11)$$

For the $(K+2)^{nd}$ arriving customer, the arrival time is $t^{(n)} + \sum_{i=2}^{K+2} a_i$ and the time of the second token's returning to the pool is $t^{(n)} + a_2 + w_2 + T$. Therefore, w^e_{K+2} and w^f_{K+2} are

$$\begin{aligned} w^e_{K+2} &= max\{0, T + w_2 - \sum_{i=3}^{K+2} a_i\} \\ w^f_{K+2} &= max\{0, w_{K+1} + s^* - a_{K+2}\} - w^e_{K+2} \end{aligned} \quad (12)$$

In general, for the l^{th} customer, $w^e_l = max\{0, T + w_{l-K} - \sum_{i=l-K+1}^{l} a_i\}$ and $w^f_l = max\{0, w_{l-1} + s^* - a_l\} - w^e_l$. The departure time from the first node is $d^f_l = \sum_{i=2}^{l} a_i + w_l + s^*$ and the departure time from the network is $d_l = \sum_{i=2}^{l} a_i + w_l + T$. For the $(l+K)^{th}$ customer, the arrival time to the external queue is $\sum_{i=2}^{l+K} a_i$ and the token's arrival time to the pool is $\sum_{i=2}^{l} a_i + w_l + T$. For the $(l+K)^{th}$ customer, we have;

$$w^e_{l+K} + w^f_{l+K} = max\{0, w_{l+K-1} + s^* - a_{l+K}\} \quad (13)$$

Now, let us consider a $GI/D/1$ queue with a service time equal to s^*. The waiting time and departure time of the j^{th} customer is;

$$w_j = max\{0, w_{j-1} + s^* - a_j\} \quad (14)$$

$$d_j = \sum_{i=2}^{j} a_i + w_j + s^* \quad (15)$$

We note that the expression for the waiting time of a customer in the network, i.e. $w_i^e + w_i^f$, is identical to (14). □

For a queueing network with constant service times, the number of tokens in the network, i.e. the size of the window, may influence where customers wait internally in the network. However, a customer's waiting time in the network is independent of the number of tokens in the network, when $K \geq K^*$ where $K^* = \lceil \frac{T}{s^*} \rceil$. This can be easily shown using the above Theorem 2. If $Ks^* \geq T$, then the waiting time of a customer is the same as that of a $GI/D/1$ queue, which is independent of K. Thus, we have the following corrolary.

Corrolary 1. The waiting time of a customer in the two-node queueing network is independent of the number of tokens K, when $K \leq K^* = \lceil \frac{T}{s^*} \rceil$.

Finally, we observe that when $K \geq N$, then $Ks^* \geq Ns^* \geq T$, and hence Theorem 2 holds. Therefore, we have the following corrolary.

Corrolary 2. If $K \geq N$, then Theorem 2 always holds.

3. Bounds on the mean waiting time

In this section, we present a lower and an upper bound of the mean waiting time in the original queueing network assuming that $Ks^* < T$. First, we consider the single token case and then the multiple token case.

3.1 The single token case

In the case where the token pool consists of one token, the original queueing network is reduced to a $GI/D/1$ queue with a service time equal to T. If the arrival process is Poisson, then we obtain an $M/D/1$ queue in general, and the mean waiting time is given by Khinchin-Pollaczek formula, i.e.,

$$W = \frac{\rho \overline{x}}{2(1-\rho)} \text{ where } \rho = \lambda \overline{x} \text{ is traffic intensity} \tag{16}$$

However, for the $GI/G/1$ queue, there is no exact expression available for the mean waiting time. Marshall [15] and Marchal [13] give the following bounds for the $GI/G/1$.

$$max\{0, \frac{\lambda^2 \sigma_B^2 + \rho(\rho-2)}{2\lambda(1-\rho)}\} \leq W_q \leq \frac{\lambda(\sigma_A^2 + \sigma_B^2)}{2(1-\rho)} \tag{17}$$

where σ_A^2 and σ_B^2 are the variances of interarrival distribution and service time distribution respectively. For the case of the constant service time, the variance σ_B^2 is zero. Therefore, the lower bound is always zero. Marchal [12] obtained the following approximation for the lower bound of the mean waiting time in a $GI/G/1$ queue;

$$\frac{\lambda^2(\sigma_A^2 + \sigma_B^2)}{2\lambda(1-\rho)} - \frac{1+\rho}{2\lambda} \tag{18}$$

One of the well-known properties of (17) and (18) is that the mean waiting time tends to the upper bound as ρ increases (see Marchal [13]). Another approximation formula for the mean waiting time in a $GI/G/1$ queue given by Kramer and

Langenbach-Belz [9] is

$$W_q = \frac{\rho^2(c_a^2 + c_s^2)}{2\lambda(1-\rho)} g(c_a^2, c_s^2, \rho) \tag{19}$$

where c_a^2 and c_s^2 are the squared coefficient of variation of the interarrival time and service times respectively, and

$$g(c_a^2, c_s^2, \rho) = \begin{cases} exp(-2(1-\rho)\frac{(1-c_a^2)^2}{3\rho(c_a^2+c_s^2)}) & \text{if } c_a^2 < 1 \\ \\ exp(-(1-\rho)\frac{(c_a^2-1)}{(c_a^2+4c_s^2)}) & \text{otherwise} \end{cases}$$

3.2 The multiple token case

Below, we obtain a lower and upper bound on the mean waiting time, assuming that $Ks^* < T$.

3.2.1 A lower bound on the mean waiting time

Let us consider the equivalent queueing network shown in Figure 3. We note that the tokens are used in the order in which they arrive at the token pool. For presentation purpose, let us number the tokens from 1 to K. Then, since service times are all constant, token i will always be behind token $(i-1)$. In view of this, every K^{th} arriving customer will use the same token. If we regard each token as a separate server, the queueing network can be represented by K queues in parallel. Each queue will consist of customers waiting to use the same token. The service time at each queue is the time it takes for a token to traverse the two nodes inside

the semaphore controlled queueing network. Obviously, this service time depends on how many other tokens are being used at the same time. In other words, the service time in a queue depends on the state of the remaining $(K-1)$ queues.

A lower bound on the mean waiting time can be easily obtained by setting the service time of each of these K queues equal to T $(= \sum_{i=1}^{N} s_i)$, i.e., independent of the state of the other queues. If the arrival process to the original queueing network is a general independent process with arrival rate λ, then the arrival process to each of the K queues is the convolution of K general arrival process with an arrival rate $\frac{\lambda}{K}$. Thus, each queue can be analyzed as a $GI \otimes GI \otimes \ldots \otimes GI/D/1$ queue, where $GI \otimes GI \otimes \ldots \otimes GI$ is the convolution of the K arrival processes, and the service time is equal to T. When the arrival process to the queueing network is Poisson with an arrival rate λ, the arrival process to each queue becomes an Erlang distribution with K phases and an arrival rate λ for each phase.

3.2.2 An upper bound on the mean waiting time

Let us consider the queueing network under study assuming that the external queue is saturated. That is, there is always at least one customer waiting in the external queue. In this case, all K tokens are continuously used. Let us consider the case where $\frac{T}{K} > s^*$. In this case, a token arriving to the first node always finds the node empty. Thus, the time it takes for a token to return to the token pool is $s^* + \overline{s} = T$ and the average departure rate for customer is $\frac{T}{K}$. Thus, when the

external queue is saturated, the average throughput is $\frac{K}{T}$. The conjecture from this is that the average throughput of the two-node queueing network lies between $\frac{1}{s^*}$ and $\frac{K}{T}$. Therefore, An upper bound on the mean waiting time can be obtained by representing a $GI/D/1$ queue with a service time equal to $max\{s^*, \frac{T}{K}\}$.

The mean waiting time in this queue can be obtained using (16) for the arrival process is Poisson. For non-Poisson arrival process, we can obtain the bounded mean waiting time by (17) or the approximation value by (19).

3.3 Stability conditions

Generally, it is known that the stability condition for the queueing network is

$$\frac{\lambda T}{K} \tag{20}$$

We note that both analytic models for the lower and upper bound satisfy this condition. In the case of the lower bound, the $GI \otimes GI ... \otimes GI/D/1$ queue is stable when $\frac{\lambda}{K}T < 1$. In the case of the upper bound, the stability condition of the $GI/D/1$ with service time $\frac{T}{K}$ is also $\lambda\frac{T}{K} < 1$.

3.4 An improved bound on the mean waiting time

In the previous section, we obtained an upper and lower bounds of the mean waiting time in the queueing network using single server queues. Below, we obtain tighter bounds by appropriately combining these single server queues.

3.4.1 An improved upper bound

During the time that the external queue is busy (i.e., the token pool is empty), we approximate the mean waiting time in the two-node queueing network by that of a $GI/D/1$ queue with service time $\frac{T}{K}$ (i.e., the upper bound). This is because, in this case, the average throughput is $\frac{K}{T}$. However, during the time that the external queue is idle, we approximate the mean waiting time in the queueing network by that of a $GI/D/1$ queue with service time s^*. This is because, in this case, an arriving customer may go straight into the first node (if there is a token in the pool upon arrival) and the average throughput of the two-node queueing network lies between $\frac{1}{s^*}$ and $\frac{K}{T}$. Let P_b and P_i be the probability that the external queue of the two-node queueing network is busy and idle respectively. We have $P_i = 1 - P_b$, and P_b is approximately taken to be equal to $\frac{\lambda T}{K}$. Thus, the mean waiting time in the queueing network is approximated by the quantity :

$$W_u = P_b W_1 + P_i W_2 \tag{21}$$

where W_1, the mean waiting time of a $GI/D/1$ queue with service time $\frac{T}{K}$ and W_2, the mean waiting time for a $GI/D/1$ queue with service time s^*.

For the Poisson arrival process, we first obtain probabilities P_b and P_i that an $M/D/1$ queue with service time $\frac{T}{K}$ is busy and idle respectively. We have $P_b = \rho$ and $P_i = 1 - \rho$, where $\rho = \frac{\lambda T}{K}$. Let W_1 and W_2 be the mean waiting time in the $M/D/1$ queue with service time $\frac{T}{K}$ and s^* respectively. The improved upper bound on the mean waiting time can be obtained by combining W_1 and W_2 as follows,

$$\begin{aligned} W_u &= P_b W_1 + P_i W_2 \\ &= \rho \frac{\lambda(\frac{T}{K})^2}{2(1 - \frac{\lambda T}{K})} + (1 - \rho) \frac{\lambda(s^*)^2}{2(1 - \lambda s^*)} \end{aligned} \tag{22}$$

For non-Poisson arrival processes, we do not have an exact expression for the mean waiting time in a $GI/D/1$ queue. In this case, we can use one of the approximations mentioned in section 3.1.

3.4.2 An improved lower bound

This lower bound in section 3.2.1 was derived by decomposing the two-node queueing network into K parallel and independent queues. Let $\rho = \frac{\lambda T}{K}$ be the probability that one of the parallel queues is busy. Then, ρ^K is the probability that all K queues are busy. When all K tokens in the two-node queueing network are used, this is approximately equivalent to all parallel queues being busy.

The improved lower bound is obtained as follows. When all tokens are being used, the mean waiting time in the two-node queueing network is approximated

by that of a $GI/D/1$ queue with service time $\frac{T}{K}$. This event is assumed to occur approximately with probability ρ^K. However, when not all tokens are used, the mean waiting time in the two-node queueing network is approximated by the lower bound given in section 3.2.1, referred to as W_3 . Thus, the improved lower bound on the mean waiting time, W_l can be obtained by combining W_1 and W_3 as follows:

$$W_l = \rho^K W_1 + (1 - \rho^K) W_3 \tag{23}$$

4. Numerical output

The lower and upper bounds given in section 3.4 were checked by comparing them against simulation estimates of the mean waiting time in the queueing network. The experiments were carried out assumming Poisson arrivals and phase-type arrival process, i.e., Erlang and hyper-exponential distributions.

Each table below gives the lower and upper bounds and the simulated mean waiting time as a function of the number of tokens ($K < K^*$). The bounds are calculated using Kramer and Langenbach-Belz [9] approximation and also Marchal's [13] bounded approach. Since these calculations are all approximated value, we have also simulated the two queueing models that are used to obtain the upper and lower bound. The obtained simulation results are also given in each table under the heading "simulated upper bound" or "simulated lower bound".

Example 1

A Poisson arrival process is assumed with an arrival rate $\lambda = \frac{1}{7}$, $s^* = 2.5$ and $\bar{s} = 28$.

number of tokens	5	6	7	8	9	10
Marchal's upper bound	23.72	9.28	5.77	4.18	3.28	2.70
K. and L.B. upper bound	18.10	5.17	2.49	1.56	1.13	0.90
simulated upper bound	18.10	5.17	2.49	1.56	1.13	0.90
sim. mean waiting time	17.45	4.25	1.80	1.06	0.80	0.70
simulated lower bound	17.11	3.99	1.30	0.48	0.20	0.10
K. and L.B. lower bound	16.60	2.78	0.52	0.09	0.01	0.01
Marchal's lower bound	11.92	1.36	0.21	0.03	0.01	0.00

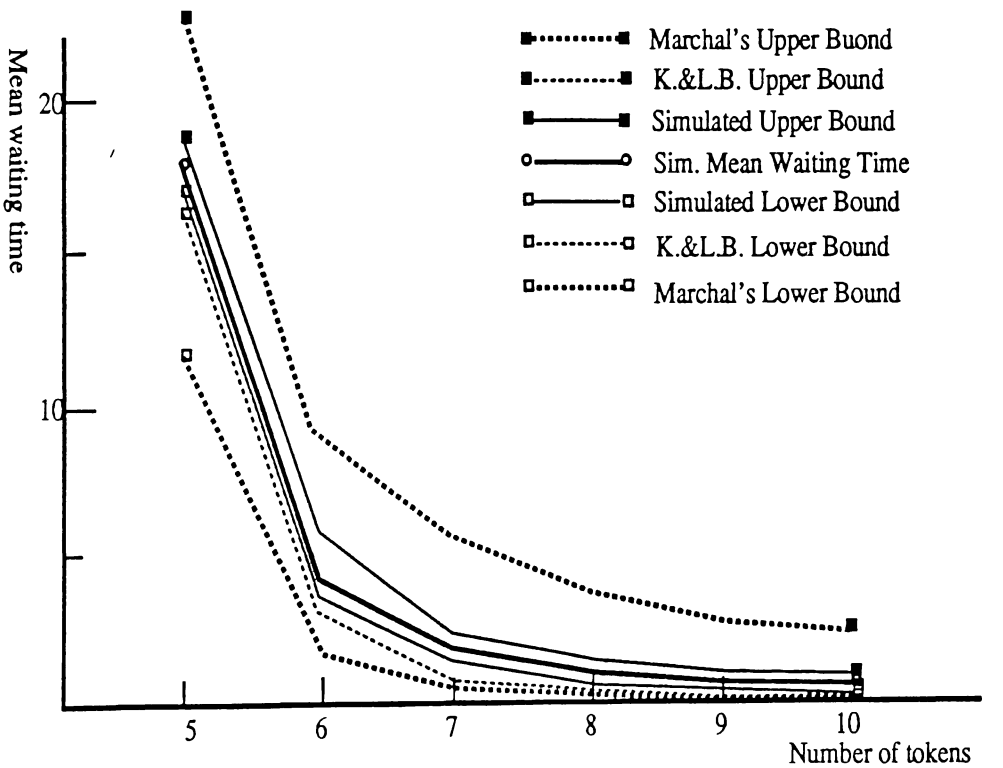

Figure 4: An improved bounds for Poisson arrival process

Example 2

The arrival process is an Erlang 2 with a phase arrival arrival rate $= \frac{1}{3.5}$

$s^* = 2.5$ and $\overline{s} = 28$.

number of tokens	5	6	7	8	9	10
Marchal's upper bound	11.86	4.64	2.89	2.09	1.64	1.35
K. and L.B upper bound	9.91	3.12	1.67	1.11	0.83	0.68
simulated upper bound	8.53	2.13	0.97	0.55	0.36	0.28
sim. mean waiting time	7.24	1.30	0.40	0.25	0.20	0.20
simulated lower bound	7.88	1.33	0.30	0.10	0.001	0.0
K. and L.B lower bound	7.64	0.77	0.07	0.01	0.00	0.00
Marchal's lower bound	5.96	0.68	0.10	0.02	0.00	0.00

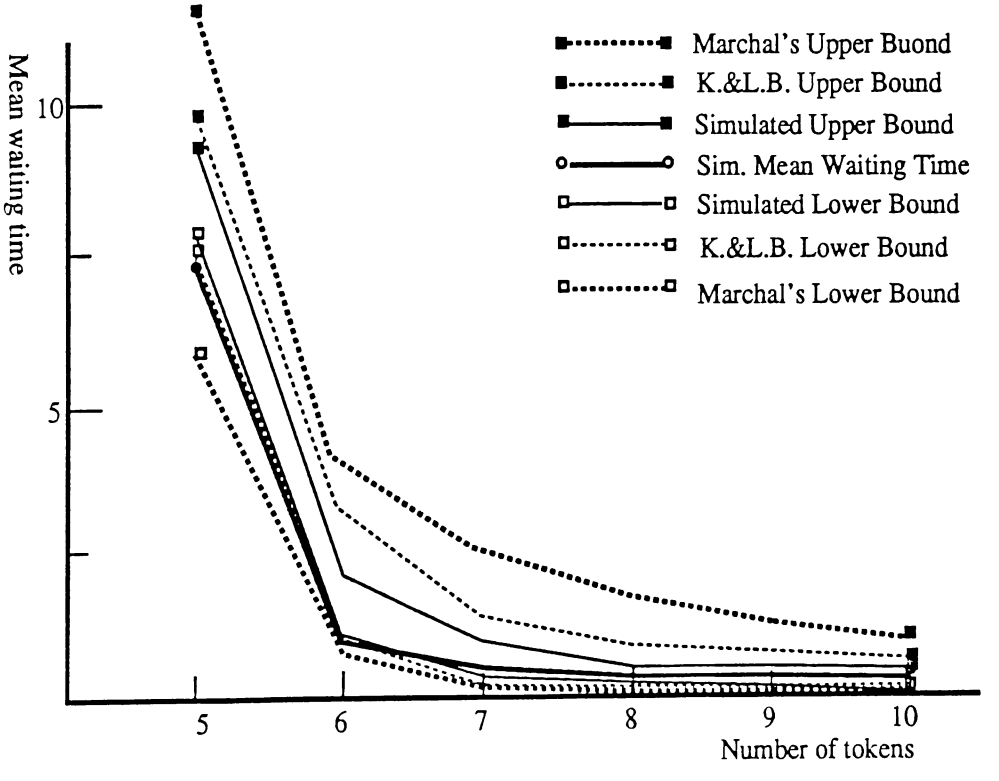

Figure 5: An improved bounds on Erlang arrival process

Example 3

The arrival process is a H_2 with $p_1 = \frac{1}{3}$, $p_2 = \frac{2}{3}$, $\lambda_1 = \frac{1}{15}$ and $\lambda_2 = \frac{1}{3}$. The squared coefficient of the variation for the arrival process $c_v^2 = 2.31$, $s^* = 2.5$ and $\overline{s} = 28$.

number of tokens	5	6	7	8	9	10
Marchal's upper bound	55.71	23.54	16.25	13.21	11.60	10.64
K. and L.B. upper bound	52.10	19.92	12.69	9.74	8.24	7.38
simulated upper bound	44.07	12.94	5.81	3.55	2.30	2.00
sim. mean waiting time	43.54	12.41	5.71	3.26	2.22	1.77
simulated lower bound	42.46	11.62	4.79	2.21	1.03	0.50
K. and L.B. lower bound	51.77	17.15	7.84	3.68	1.63	0.68
Marchal's lower bound	42.94	3.45	0.59	0.09	0.02	0.002

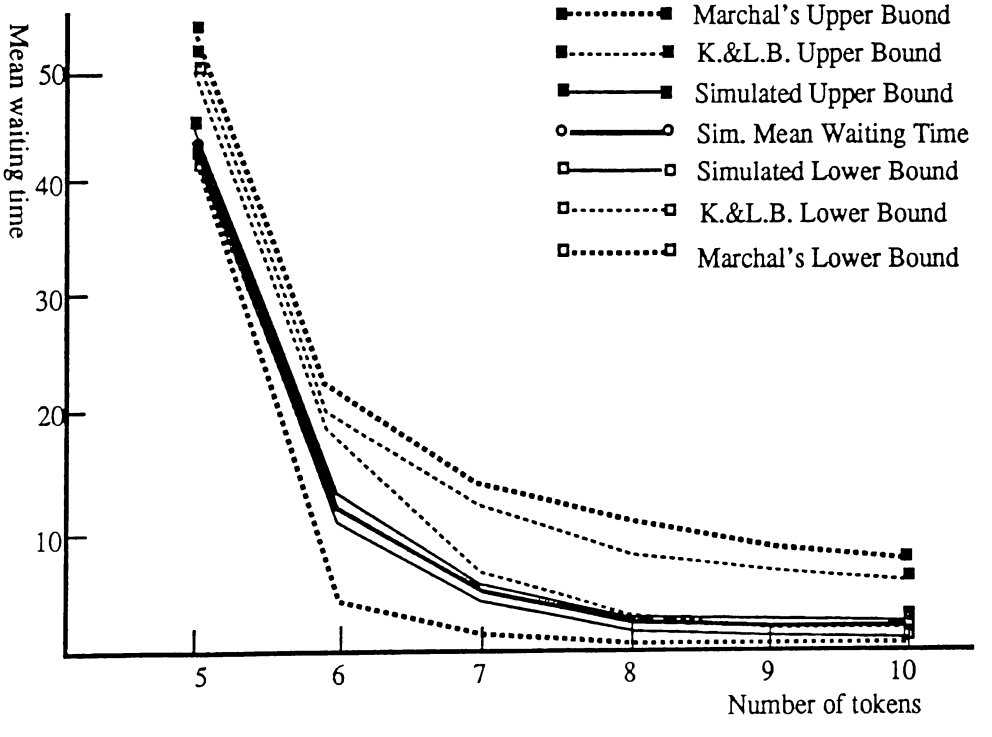

Figure 6: An improved bounds on hyper exponential process

5. Conclusions

We consider an open tandem queueing network with population constraint and constant service times.

It is shown that this queueing network can be transformed into a simple queueing network involving only two nodes. Using this simple queueing network, we obtain an upper and lower bound on the mean waiting time. These bounds can be easily calculated. Simulation experiments showed that the lower bound is a very good approximation when the number of tokens is small. The upper bound also gives a good approximation as the number of tokens increases.

References

[1] T. Altiok and Z. Kao, "Bounds for throuput in production/inventory systems in series with deterministic processing times", *Dept. of IE Rutgers Univ.,* May 1987.

[2] B. Avi-Itzhak, "A sequence of service stations with arbitrary input and regular service times", *Mangement Science,* 11(1965) 565-571.

[3] Y. Dallery, "Approximate analysis of general open queueing networks with restricted capacity", *Technical Report LAG,* N87-08, 1987, revised Feb. 1989.

[4] S. Fdida, H. Perros, and A. Wilk, "Semaphore queues : Modeling multi-layered window flow control mechanisms ", *IEEE Trans. Comm.,* 38, pp.309-317, 1990.

[5] H.D. Friedman, "Reduction methods for Tandem queueing system." *Operations Research,* V-13, 1965.

[6] P. L. Gall, "Packetized queueing networks and window flow control." *CNET,* Issy-les-Moulineaux, France.

[7] J.S. Kaufman and Y.J. Wang, "Approximate analysis of a simultaneous resource possesion problem", *ICCC,* pp. 199-206, 1988.

[8] J.F.C. Kingman, "Some inequalitiesfor the queue GI/G/1", *Biometrika,* v-49. pp. 315-324, 1962.

[9] Kramer and Langenbach-Belz, "Approximate formulae for the delay in the queueing system GI/G/1." *Eight Int. Tele. Con.,* Melbourne, 235/1-8

[10] S.S. Lam, "Queueing networks with population size constraint," *IBM J. RD,* v-197, pp. 370-378.

[11] Lindley, "Theory of queue with single server", *Proc. Camb. Phil. Soc.,* 48 pp. 277-289, 1952.

[12] W.G. Marchal, "A modified Erlang approach to approximating GI/G/1 queues", *J.Appl.Prob.,* 13, pp. 118-126, 1976.

[13] W.G. Marchal, "Some simpler bounds on the mean queueing time", *Operations Research,* v-26, No 6, pp. 1083-1088, 1978.

[14] R. Marie, "An approximate analytical method for general queueing networks", *IEEE Trans. on Software Engineering,* SE-5, n5, pp. 530-538, 1979.

[15] K.T. Marshall, "Some inequalities in queueing", *Operations Research,* v-16, pp. 651-665, 1968.

[16] G.F. Newell. "Approximate behavior of tandem queues", *Lecture note in Economics and Mathematical system,* pp. 171 (Springer, Berlin, 1979).

[17] H.G. Perros, "Queueing Networks with Blocking ", *Performance Evaluation Review,* 12, pp. 8-12, 1984.

[18] H.G. Perros, Y.Dallery and G. Pujolle, "Analysis of a queueing network model with class dependent window flow control", *IEEE Infocom,* pp. 968-977, 1992.

[19] M. Reiser, "Admission delays on virtual ruotes with window flow control", in *Proc. Perform. Data Commun. Syst. Appl.,* Pujolle, Ed., North-Holland, pp. 67-76, 1981.

[20] M. Shalmon and M. Kaplan, "A tandem network of queues with deterministic service and intermediate arrivals", *Operations Research,* v-32, n4, pp. 753-773, 1984.

[21] G.W. Shapiro and H.G. Perros, "Nested sliding window protocols with packet fragmentation", *CCSP TR-89/15,* North Carolina State University.

[22] C. Ziegler and D.L. Schilling, "Delay decomposition at a single server queue with constant service time and multiple inputs", *IEEE Trans. on Commun.,* v-26, No 2, pp. 290-295, 1978.