An Approximate 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 present 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.

This queueing network is transformed into an equivalent simpler queueing network involving only two nodes. Using this simple queueing network, we obtain upper and lower bounds for the mean waiting time and for the variance of the interdeparture time. Approximations for the mean waiting time and the variance of the interdeparture time are then constructed by appropriately combining their upper and lower bounds. Validations against simulation showed that these approximations have a low relative error.

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 constraint have been used to model flow control mechanism in data communication systems, automatic assembly lines with a fixed number of palets, and semaphore controlled software in an operating system.

Queueing networks with population constraints do not have a closed form solution. As a result, several approximate solutions have been proposed in the literature. Fdida, Perros and Wilk [5] 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 has a population constraint. They used this type of queueing network to model nested sliding window flow control mechanisms. Dallery [4] analyzed a single class open queueing network with population constraint under Coxian service times by transforming it into an equivalent closed queueing network. Shapiro and Perros [18] presented a hierarchical method for analyzing nested sliding window flow control mechanisms with packet fragmentation and reassembly. Perros, Dallery and Pujolle [15] extended the approach proposed in [4] in order to analyze open multiclass queueing networks with class dependent population constraints. Other semaphore controlled queueing models have also been considered in the literature. 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 [8] 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 [6] 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 [20] and Gall [7] obtained the delay in a queueing network with Poisson arrivals and constant service times assuming infinite capacity queue. Shalmon and Kaplan [17] 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. Finally, Newell [14] analyzed a tandem network with constant service times, assuming that the first node is saturated.

In this paper, we analyze an open tandem queueing network with population constraint assuming constant service times. In particular, we obtain approximately the mean waiting time in the queueing network and the variance of the interdeparture time. To the best of our knowledge, this model has not as yet been studied in the open literature. This paper builds on an earlier paper by Rhee and Perros [16]. It was shown in [16] that this queueing network can be transformed into a simple queueing network involving only two nodes. Using this simple queueing network, an upper and lower bound on the mean waiting time were constructed. These bounds can be easily calculated. Simulation experiments showed that the lower bound is a very good approximation to the mean waiting time when the number of tokens is small. The upper bound also gives a good approximation as the number of tokens increases.

The paper is organized as follows. In section 2, we present the open queueing network under study. In section 3, we give an upper and lower bound of the mean waiting time in the queueing network, and then we obtain an approximation of the mean waiting time by appropriately weighing these two bounds. In section 4, we obtain approximately the variance of the interdeparture time following the same approach as in section 3. That is, we first obtain an upper and a lower bound

on the variance of the interdeparture time. Then, we construct an approximation to this variance by weighing these two bounds. Finally, conclusions are given in section 5.

We note that, in this paper, we interpret the waiting time of a customer as the total time a customer spends queueing up in the queueing network, rather than the time it takes to traverse the queueing network which also includes service times.

2 The Queueing Network Model under Study

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 an arbitrary general distribution with rate λ, and the service time at each node i 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.

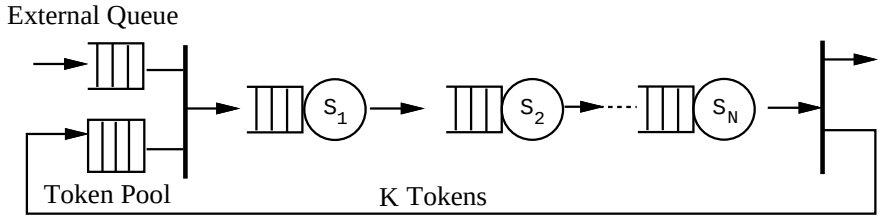

Figure 1: An open queueing network with constant service times

The semaphore is a mechanism that 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 without delay. 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. In particular, let us rearrange the nodes of the open tandem queueing network so that the node with the longest service time is placed at the beginning of the queueing network. Then, a customer's waiting time is the same in both the rearranged queueing network and in the original queueing network (for a proof see in [16]). Since there is no queueing after the first node in the rearranged queueing network, 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 by a simpler two-node queueing network as shown in Figure 2.

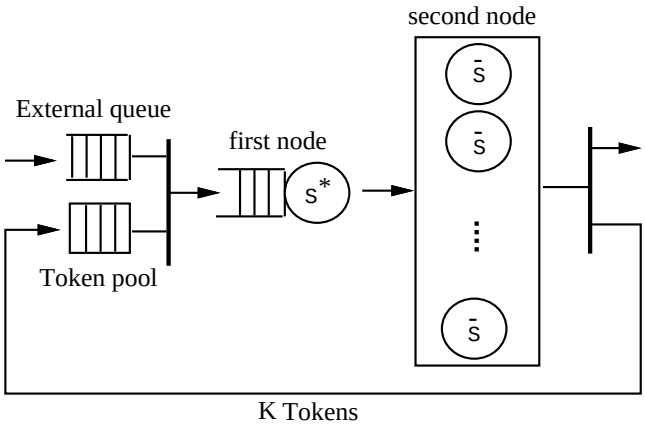

Figure 2: The two-node queueing network with constant service times

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. In this paper, we focus on the simpler two-node queueing network. The number of tokens in the network affects a customer's waiting time. In the following theorem, we prove that a customer's waiting time is monotonically decreasing as the number

of tokens increases. (We note that the same theorem has been proved for general service times for stochastic decision free petri nets by Baccelli and Liu [3]. Below we include the proof to this theorem since it is directly related to constant service times, and some of arguments are used in the rest of the paper.)

Theorem 1. Let us consider the semaphore controlled queueing network with constant service times. Let $w_i(K)$ be the waiting time of the i^{th} customer in the queueing network with K tokens. Then

$$w_i(K) \geq w_i(K+1) \qquad \text{for all } i \tag{1}$$

Proof. For the first K arriving customers, $w_i(K) = w_i(K+1)$ for $i \leq K$, since an arriving customer always finds a token in the token pool. Letting a_i be the interarrival time between the $(i-1)^{st}$ and i^{th} customer, we have

$$w_i(K) = max\{0, w_{i-1}(K) + s^* - a_i\} \; 2 \leq i \leq k$$

Let $T = \sum_{i=1}^{N} s_i$. Then, for the $(K+1)^{st}$ arriving customer, we have

$$w_{K+1}(K) \;\; = \;\; max\{0, max\{T, \sum_{j=1}^{K} a_j + w_K(K) + s^*\} - \sum_{j=1}^{K+1} a_j\} \tag{2}$$

and

$$w_{K+1}(K+1) \;\; = \;\; max\{0, w_K(K+1) + s^* - a_{K+1}\} \tag{3}$$

Since $w_K(K) = w_K(K+1)$, we have that $w_{K+1}(K) \geq w_{K+1}(K+1)$. At this point it is reasonable to conjecture that

$$w_i(K) \geq w_i(K+1) \text{ for all } i \geq K \tag{4}$$

We prove that this is, in fact, the case by mathematical induction. Expression(4) clearly holds for the first $K+1$ values of the index i. Assuming that it holds for $i-1$, we show below that it also holds for i. We have

$$w_i(K) \;\; = \;\; max\{0, max\{T, \sum_{j=i-K+1}^{i} a_j + w_{i-K}(K) + s^*\} - \sum_{j=1}^{i+1} a_j\} \tag{5}$$

and

$$w_i(K+1) \;=\; max\{0, w_{i-K}(K+1) + s^* - a_{i+1}\} \tag{6}$$

Since $w_{i-K}(K) \geq w_{i-K}(K+1)$, we conclude that $w_i(K) \geq w_i(K+1)$. $\Box$

We note that after a given number of tokens K^*, a customer's waiting time does not change even though the number of tokens is increased. We showed that $K^* = \lceil \frac{T}{s^*} \rceil$ in [16]. If $Ks^* \geq T$, then the waiting time in the queueing network is the same as that of a $G/D/1$ queue with the same arrival process and service time s^*. That is, it is independent of K. We also observe that when $K \geq N$, then $Ks^* \geq Ns^* \geq T$.

3 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 $Ks^* < T$. By appropriately weighing these two bounds, we obtain an approximation for the mean waiting time. We note that these two bounds were reported in an earlier paper by Rhee and Perros [16]. For presentation purposes, we describe these two bounds below in section 3.1 and 3.2.

In the case where the token pool consists of one token, since the network allows only one customer, there is no waiting after the external queue. The original queueing network becomes a $G/D/1$ queue with a service time equal to T. If the arrival process to the queueing network is Poisson, then we obtain an $M/D/1$ queue, and the mean waiting time W is given by the Khinchin-Pollaczek formula, i.e.,

$$W = \frac{\rho T}{2(1-\rho)} \tag{7}$$

where $\rho = \lambda T$ is traffic intensity. However, for the $G/G/1$ queue, there is no exact expression available for the mean waiting time. Marshall [11] and Marchal [13] give the following bounds for the $G/G/1$ queue.

$$max\{0, \frac{\lambda^2\sigma_B^2 + \rho(\rho-2)}{2\lambda(1-\rho)}\} \leq W \leq \frac{\lambda(\sigma_A^2 + \sigma_B^2)}{2(1-\rho)} \tag{8}$$

where σ_A^2 and σ_B^2 is the variance of the interarrival time and service time respectively. For the case of constant service times, the variance σ_B^2 is zero. Therefore, the lower bound is always zero since $\rho < 1$. Marchal [12] obtained the following approximation for the lower bound of the mean waiting time in a $G/G/1$ queue :

$$\frac{\lambda^2(\sigma_A^2 + \sigma_B^2)}{2\lambda(1-\rho)} - \frac{1+\rho}{2\lambda} \tag{9}$$

Further, the mean waiting time in a $G/D/1$ queue can be approximated by the following expression due to Kramer and Langenbach-Belz [9] :

$$W = \frac{\rho^2(c_a^2 + c_s^2)}{2\lambda(1-\rho)} g(c_a^2, c_s^2, \rho) \tag{10}$$

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

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

3.1 A Lower Bound

Let us consider the equivalent queueing network shown in Figure 2. We note that tokens are used in the order in which they arrive at the token pool. For presentation purposes, let us number them 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 arrival process with arrival rate λ, then the arrival process to each of the K queues is the convolution of K such general arrival processes. Thus, each queue can be analyzed as a $G \otimes G \otimes \ldots \otimes G/D/1$ queue, where $G \otimes G \otimes \ldots \otimes G$ 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 with a parameter λ for each phase. For Poisson and non-Poisson arrivals, the mean waiting time is calculated using (8) or (10).

3.2 An Upper Bound

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 $\frac{T}{K} > s^*$. Since the external queue is always saturated, sooner or later there will be no token left in the token pool. The interdeparture time from the first node is larger than or equal to s^*, which means that the interarrival time of a token to the pool is larger than or equal to s^*. Thus, a token arriving to the first node always finds the node empty. The time it takes for a token to return to the token pool is $s^* + \overline{s} = T$ and the average interdeparture time of a customer is $\frac{T}{K}$. Thus, when the external queue is saturated, the throughput is $\frac{K}{T}$. We conjecture from this, that the 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 $G/D/1$ queue with a service time equal to $max\{s^*, \frac{T}{K}\}$. The mean waiting time in this queue can be obtained using (7) if the arrival process is Poisson. For a non-Poisson arrival process, we use (8) or (10).

3.3 The True Boundness of the Upper and Lower Bound

In this subsection, we show that the above lower and upper bounds are true bounds of the mean waiting time in the original tandem queueing network.

Theorem 2. Let w_i^l and w_i^u be the waiting time of the i^{th} arriving customer in the lower and upper bound model respectively. Let w_i be also the waiting time of the i^{th} arriving customer in the original semaphore controlled queueing network. Then, the waiting time w_i, is always bounded by the lower and upper bounds, that is,

$$w_i^l \leq w_i \leq w_i^u \text{ for all } i \tag{11}$$

Proof. Let us consider the beginning of the first busy period of the semaphore controlled queueing network. For the first K arriving customers, we have

$$w_i^l = 0 \tag{12}$$

$$w_i = max\{0, w_{i-1} + s^* - a_i\} \tag{13}$$

$$w_i^u = max\{0, w_{i-1}^u + \frac{T}{K} - a_i\} \tag{14}$$

where a_i is the interarrival time between the $(i-1)^{st}$ and i^{th} customer.
Since $Ks^* \leq T$, we can prove easily that $w_i^l \leq w_i \leq w_i^u$ where i takes the values $1 \leq i \leq K$. For the $K+1^{st}$ arriving customer, we have

$$w_{K+1}^l = max\{0, T - \sum_{i=1}^{K+1} a_i\} \tag{15}$$

$$w_{K+1} = max\{0, max\{T, \sum_{i=1}^{K} a_i + w_K + s^*\} - \sum_{i=1}^{K+1} a_i\} \tag{16}$$

$$w_{K+1}^u = max\{0, w_K^u + \frac{T}{K} - a_{K+1}\} \tag{17}$$

If $T > \sum_{i=1}^{K} a_i + w_K + s^*$, then $w_{K+1}^l = w_{K+1}$. Otherwise $w_{K+1}^l < w_{K+1}$. In order to compare w_{K+1} and w_{K+1}^u, we re-write w_{K+1}^u as follows :

$$w_{K+1}^u = max\{0, w_K^u + \frac{T}{K} - a_{K+1}\}$$

$$
\begin{aligned}
&= max\{0, w^u_{K-1} + \frac{2T}{K} - \sum_{i=K}^{K+1} a_i\} \text{ if } w^u_{K-1} + \tfrac{T}{K} - a_K > 0 \\
&\vdots \\
&= max\{0, w^u_j + \frac{(K-j+1)T}{K} - \sum_{i=j+1}^{K+1} a_i\} \text{ if } w^u_j + \tfrac{T}{K} - a_{j+1} > 0 \\
&\vdots \\
&= max\{0, w^u_1 + \frac{KT}{K} - \sum_{i=2}^{K+1} a_i\} \text{ if } w^u_1 + \tfrac{T}{K} - a_2 > 0 \qquad (18)
\end{aligned}
$$

From (16), if $T \geq \sum_{i=1}^{K} a_i + w_K + s^*$, then $w_{K+1} = max\{0, T - \sum_{i=2}^{K+1} a_i\}$. Otherwise, $w_{K+1} = max\{0, w_K + s^* - a_{K+1}\}$. However, from (17), we have that if $w^u_{K-1} + \frac{T}{K} - a_K > 0$, then $w^u_{K+1} = max\{0, w^u_K + \frac{T}{K} - a_{K+1}\}$, and if both $w^u_{K-1} + \frac{T}{K} - a_K > 0$ and $w^u_{K-2} + \frac{T}{K} - a_{K-1} > 0$ are satisfied, then $w^u_{K+1} = max\{0, w^u_{K-1} + \frac{T}{K} - a_K\}$. Finally, we have that $w^u_{K+1} = max\{0, T - \sum_{i=2}^{K+1} a_i\}$, when for all $1 \leq i \leq K$ $w^u_i + \frac{T}{K} - a_{i+1} > 0$. Hence, $w^u_{K+1} \geq max\{0, T - \sum_{i=2}^{K+1} a_i\}$. Therefore, $w_{K+1} \leq w^u_{K+1}$.

In general, w^l_j, w_j and w^u_j, where $j \gg K$, are as follows :

$$
\begin{aligned}
w^l_j &= max\{0, w^l_{j-K} + T - \sum_{i=j-K+1}^{j} a_i\} \qquad (19) \\
w_j &= max\{0, max\{w_{j-K} + T, \sum_{i=1}^{j-1} a_i + w_{j-1} + s^*\} \\
&\quad - \sum_{i=j-K+1}^{j} a_i\} \qquad (20) \\
w^u_j &= max\{0, w^u_{j-1} + \frac{T}{K} - a_j\} \\
&\geq max\{0, w^u_{j-K} + T - \sum_{i=j-K+1}^{j} a_j\} \qquad (21)
\end{aligned}
$$

Since we already know that $w^l_{j-K} \leq w_{j-K} \leq w^u_{j-K}$, we can prove that $w^l_i \leq w_i \leq w^u_i$ for all i by applying (19), (20) and (21). □

3.4 An Approximation for the Mean Waiting Time

In this section, we obtain an approximation for the mean waiting time in the semaphore controlled queueing network by appropriately weighing up the lower and upper bounds presented in section 3.1 and 3.2. From experimental evidence we have observed that the exact mean waiting time tends to the lower bound as $\frac{Ks^*}{T}$ becomes small (i.e. ≤ 0.5), and it tends to the upper bound as $\frac{Ks^*}{T}$ goes to 1, as shown in Figure 3.

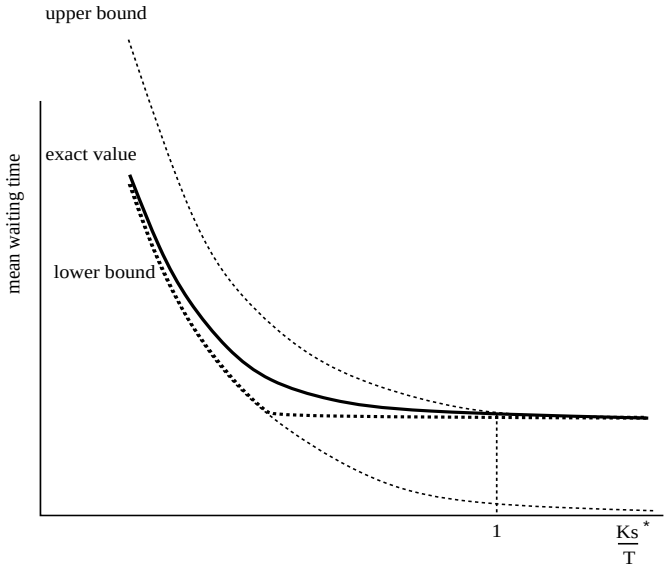

Figure 3: The exact mean waiting time and its bounds.

The upper bound converges to a certain point as the number of tokens K increases. From Theorem 2 in [16], it can be easily shown that this point is the mean waiting time in a $G/D/1$ queue with service time s^*. Since the lower bound tends to zero as the number of tokens K increases, a better lower bound can be constructed by taking the maximum of the mean waiting time in a $G/D/1$ queue with service time s^* and the mean waiting time in a $G \otimes G \otimes \ldots \otimes G/D/1$ queue with service time T. This lower bound is shown in Figure 3 with a thick dotted line. The exact mean value, obtained by simulation, is drawn in Figure 3 as a continuous thick line. We have also observed that the coefficient of variation of

the interarrival times affects how close the exact mean waiting time is to either the lower bound or the upper bound.

Let us express the distance from the exact mean waiting time to the lower bound by the following expression :

$$d = \frac{\text{exact mean waiting time } - \text{ lower bound}}{\text{upper bound } - \text{ lower bound}} \tag{22}$$

We shall refer to d as the normalized distance. We have that $0 \leq d \leq 1$. Figure 4 shows how this normalized distance d varies as a function of $\frac{Ks^*}{T}$ for different values of c_a^2, the squared coefficient of variation of the interarrival times. (These results were calculated by estimating the exact mean waiting time by simulation.) For a given value of c_a^2, d behaves in the same way as the exact mean waiting time in Figure 3. We note, however, that d tends to become linear as c_a^2 increases. Based on the above empirical observations, we constructed an approximation to the mean waiting time W_a, by combining appropriately the upper and lower bounds. In particular,

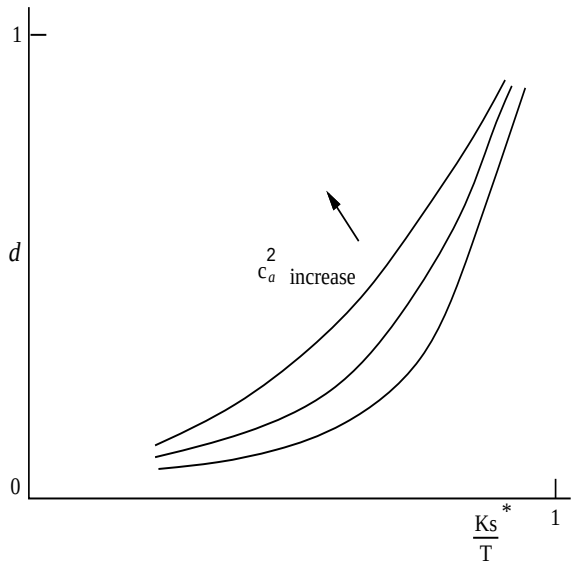

Figure 4: Normalized distance vs. c_a^2 and $\frac{Ks^*}{T}$

$$W_a \;=\; W_l + f(c_a^2, s^*, T, K)(W_u - W_l), \tag{23}$$

where W_l and W_u are the mean waiting time of the lower and upper bound respectively, and

$$f(c_a^2, s^*, T, K) \quad = \quad (\frac{K s^*}{T})^{(1+\frac{1}{c_a^2})}. \tag{24}$$

Function $f(c_a^2, s^*, T, K)$ is chosen so that it behaves in a similar fashion as the normalized distance d.

3.5 Validation

The accuracy of the approximation described in section 3.4, was checked by comparing it against simulation estimates of the mean waiting time. The experiments were carried out assuming Poisson and phase-type arrivals, i.e. Erlang and hyper-exponential distributions. Each table below gives results on the mean waiting time as a function of the number of tokens ($K < K^*$). In particular, it gives a) the simulated mean waiting time, b) approximate results obtained by simulating the lower and upper bound queueing models (referred to as simulated approximation), c) approximate results obtained using Marchal's expression (8) for the calculation of the upper and lower bounds (referred to as Marchal's approximation), and d) approximate results obtained using Kramer and Langenbach-Belz's expression (10) for the calculation of the upper and lower bounds (referred to as K. and L.B. approximation). The approximate results were calculated using a slightly different upper and lower bounds. These bounds, reported as the improved bounds in Rhee and Perros [16], are tighter than those described in section 3.1 and 3.2. When the approximation results are calculated using the bounds given in section 3.1 and 3.2, the average relative error increases by 1.5%. We note that the approximation based on Kramer and Langenbach-Belz's expression gives better results than the approximation based on Marchal's expression.

Example 1 : Poisson arrivals with $\lambda = \frac{1}{7}$, $s^* = 2.5$ and $\overline{s} = 28$.

number of tokens	5	6	7	8	9	10
Marchal's approximation	14.51	3.75	3.23	2.93	2.69	2.46
simulated approximation	17.58	4.25	1.69	1.07	0.93	0.83
sim. mean waiting time	17.45	4.25	1.80	1.06	0.80	0.70
K. and L.B. approximation	17.82	3.53	1.30	1.07	0.93	0.83

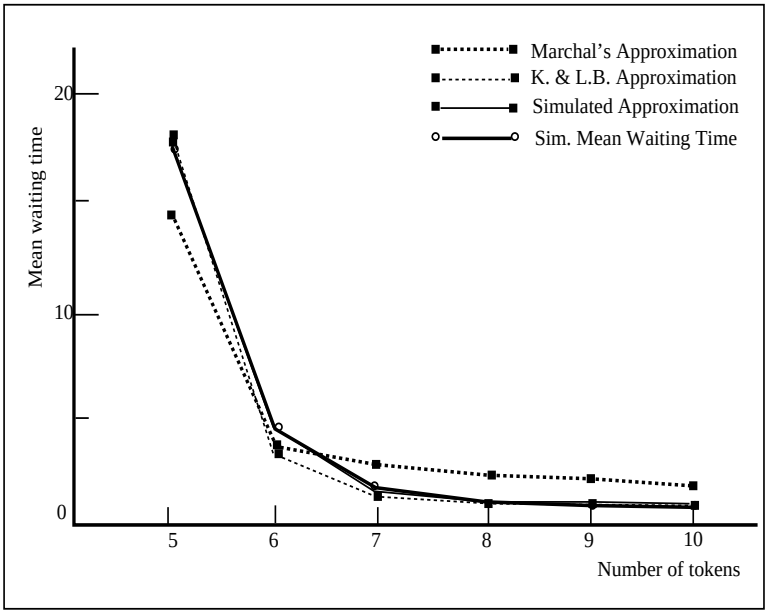

Figure 5: An approximation for Poisson($\frac{1}{7}$,2.5,28)

Example 2 : Erlang 2 phases with a phase arrival rate = $\frac{1}{3.5}$, $s^* = 2.5$ and $\overline{s} = 28$.

number of tokens	5	6	7	8	9	10
Marchal's approximation	6.37	1.59	1.50	1.43	1.36	1.27
simulated approximation	7.32	1.31	0.42	0.29	0.25	0.22
sim. mean waiting time	7.24	1.30	0.40	0.25	0.20	0.20
K. and L.B. approximation	7.34	0.85	0.34	0.30	0.27	0.23

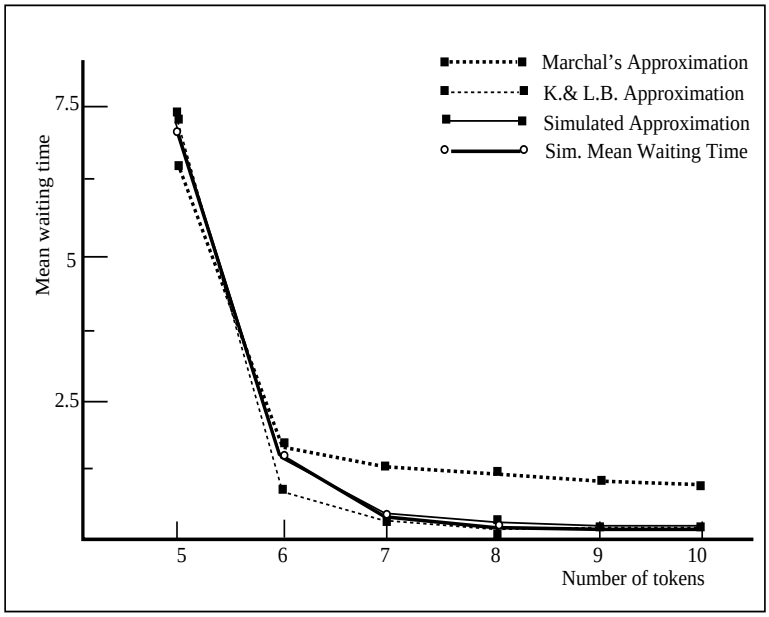

Figure 6: An approximation for Erlang(2,$\frac{1}{3.5}$,2.5,28)

Example 3 : hyper-exponential arrivals 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 $c_a^2 = 2.31$, $s^* = 2.5$ and $\overline{s} = 28$.

number of tokens	5	6	7	8	9	10
Marchal's approximation	46.49	14.74	12.68	11.64	10.94	10.42
simulated approximation	43.04	12.19	5.25	2.95	2.10	1.90
sim. mean waiting time	43.54	12.41	5.71	3.26	2.22	1.77
K. and L.B. approximation	44.58	12.58	5.36	2.61	1.58	1.37

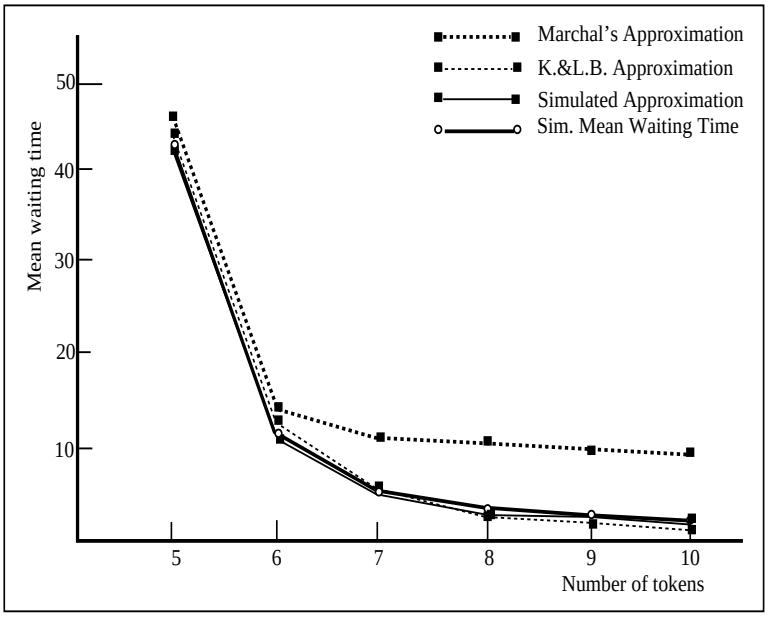

Figure 7: An approximation for Hyper($\frac{1}{3}$,$\frac{2}{3}$,$\frac{1}{15}$,$\frac{1}{3}$,2.5,28,2.31)

4 The Variance of The Departure Process

In this section, we first obtain a lower and upper bound of the variance of the inter-departure time from the semaphore controlled queueing network. Subsequently, we construct an approximation to the variance by appropriately combining these two bounds. The accuracy of the approximation is validated through simulation.

4.1 Some Basic Properties

Define a_{i0} and a_{i1} to be the interarrival time of the i^{th} customer to the external queue and to the first node. Also, let w_i^e be the waiting time of the i^{th} customer in the external queue. In Rhee and Perros [16], it was shown that (equation (6))

$$a_{i1} = a_{i0} + w_i^e - w_{i-1}^e. \tag{25}$$

By taking expectations, we have

$$E[a_{i1}] = E[a_{i0}] + E[w_i^e] - E[w_{i-1}^e]. \tag{26}$$

In the steady state, since the queueing network is stable, we have that $E[w_i^e] = E[w_{i-1}^e]$. Therefore, $E[a_{i1}] = E[a_{i0}]$. That is, as anticipated, the mean interarrival time to the external queue is equal to the mean interarrival time to the first node. Now, let us consider the variance of the interarrival time to the first node. From (25), we have

$$V[a_{i1}] = V[a_{i0}] + 2\{cov[a_{i1}, w_i^e] - cov[a_{i0}, w_{i-1}^e]\} \tag{27}$$

Since the covariance between a_{i0} and w_{i-1}^e is zero, we have

$$V[a_{i1}] = V[a_{i0}] + 2cov[a_{i1}, w_i^e] \tag{28}$$

Further, if a_{i1} and w_i^e are positively correlated. This can be shown by noting that if a_{i1} increases by a small positive value ε, then w_i^e also increases by ε, and vice versa. Thus, we have $cov[a_{i1}, w_i^e] \geq 0$ and

$$V[a_{i1}] \geq V[a_{i0}] \tag{29}$$

We note that the variance of the interdeparture time from the queueing network is the same as that from the first node. This is because, the second node is an infinite server queue with constant service time. For the case $Ks^* \geq T$, the interdeparture time from the queueing network is the same with that of a $G/D/1$ queue with service time s^*.

Finally, we note that the variance of the interdeparture time τ of a $G/D/1$ queue is given by the well-known expression,

$$V[\tau] = V[a] - 2(\frac{1}{\lambda} - s)E[w], \tag{30}$$

where λ is the arrival rate, s is the service time, a is the interarrival time and w is the waiting time.

4.2 Bounds on the Variance of the Interdeparture Time

We construct an upper and lower bound of the variance of the interdeparture time using the lower and upper bound models for the mean waiting time, presented in sections 3.1 and 3.2.

4.2.1 A Lower bound

The upper bound for the mean waiting time was obtained using a $G/D/1$ queue with a service time $\frac{T}{K}$. This queueing is equivalent to the two-node queueing network shown in case 1 of Figure 8. That is, the waiting time of the i^{th} customer is the same in both queueing systems. This can be easily shown by recalling that if $Ks^* \geq T$, then the two-node queueing network is equivalent to a $G/D/1$ queue with a service time s^* (see Theorem 2 in [16]). For this case, here, we have that $s^* = \frac{T}{K}$, and thus $(\frac{T}{K})K \geq T$.

The variance of the interdeparture time from the queueing network in case 1 of Figure 8, is —s than or equal to the variance of the interdeparture time of the two-node queueing network under study, shown in case 2, Figure 8. This can be shown intuitively as follows. From Theorem 2 in section 3.3, we have that the i^{th} customer's waiting time in case 1 is larger than that in case 2. Since the

waiting time for each customer is larger, the length of the busy period in case 1 is longer than the busy period in case 2. In view of this, one can argue that the interdeparture times in case 1 are more regular than those in case 2. Therefore, the variance of the interdeparture time in case 1 is less than the variance of the interdeparture time in the two-node queueing network under study.

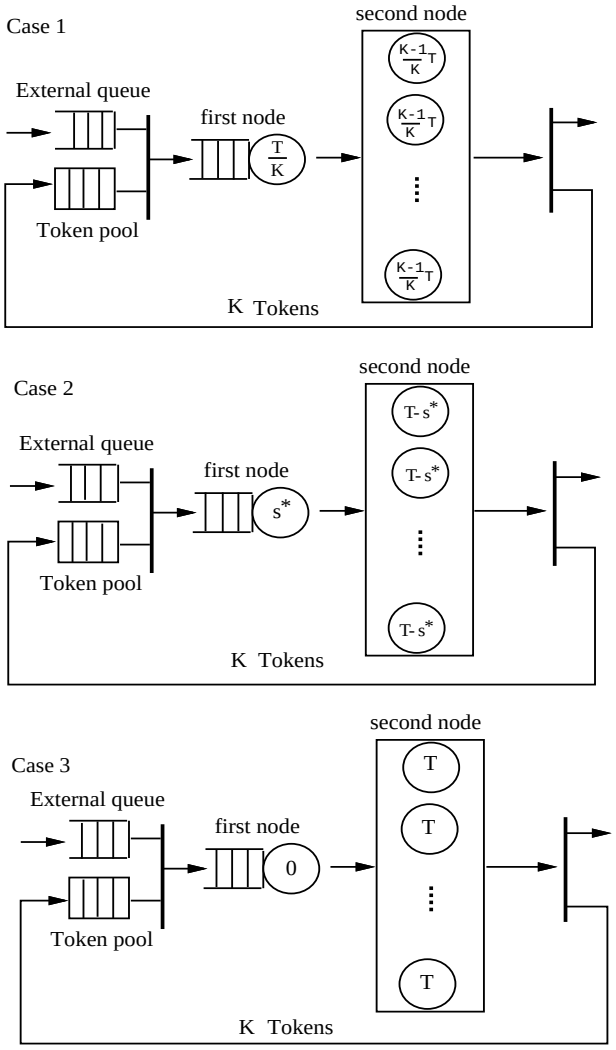

Figure 8: Two node queueing networks with constant service times

4.2.2 Upper bound

The lower bound on the mean waiting time was obtained using a queueing model consisting of K queues in parallel as shown in Figure 9. The service time at each queue is equal to T. The arrival process is cyclic, so that every K^{th} customer joins the same queue. For the purpose of obtaining the lower bound of the mean waiting time it sufficed to study one of these K queues.

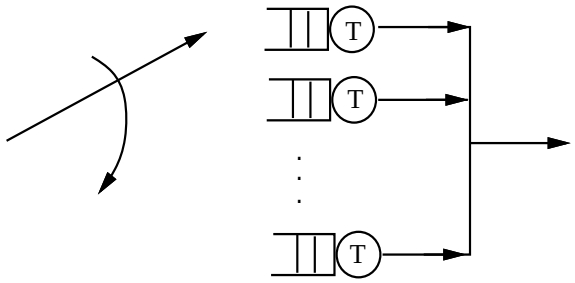

Figure 9: K queues in parallel

Below, we show that the variance of the interdeparture time of the superposition of the departure processes from the K queues of the queueing network in Figure 9, is an upper bound of the variance of the interdeparture time of the two-node queueing network under study. We first show that this superposition can be easily obtained.

Theorem 3. The superposition of the departure processes from the K queues of the queueing system shown in Figure 9, is identical to that of a single queue served by K parallel servers as shown Figure 10. The service time at each server is T and the arrival process is the same as in the queueing system in Figure 9, i.e., identical to the arrival process in the two-node queueing network under study.

Proof. Let us consider the beginning of a busy period. For the first K arriving customers to the queueing system shown in Figure 9, the departure time of the j^{th} customer from the queueing system d_j, is

$$d_j = \sum_{i=1}^{j} a_i + w_j + T$$

$$= \sum_{i=1}^{j} a_i + T \tag{31}$$

After the K^{th} arrival, the departure time of the j^{th} customer d_j, $j \geq K+1$, is

$$d_j = \sum_{i=1}^{j} a_i + w_j + T, \tag{32}$$

where $w_j = max\{0, T + w_{j-K} - \sum_{i=j-K+1}^{j} a_i\}$.

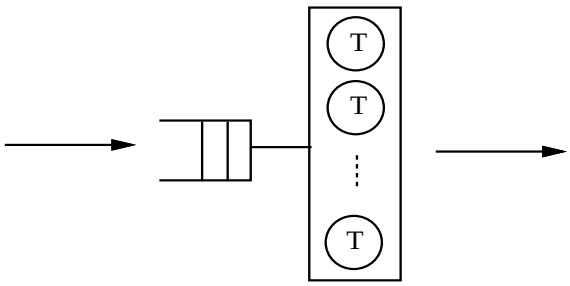

Figure 10: A single node with K servers

Now, let us consider the queueing system in Figure 10. For the first K arriving customers, we have the same departure times as above. For the j^{th} arriving customer ($j \geq K+1$), its arriving time is $\sum_{i=1}^{j} a_i$, and the earlist time it starts service is

$$d_{j-K} = \sum_{i=1}^{j-K} a_i + w_{j-K} + T. \tag{33}$$

Hence, d_j for $j \geq K+1$ is

$$d_j = \sum_{i=1}^{j} a_i + w_j + T, \tag{34}$$

where $w_j = max\{0, \sum_{i=1}^{j-K} a_i + w_{j-K} + T - \sum_{i=1}^{j} a_i\}$, which is equal to $max\{0, T + w_{j-K} - \sum_{i=j-K+1}^{j} a_i\}$. Therefore, the two queueing models are equivalent as far as a customer's departure time is concerned. □

The variance of the interdeparture time of the $G/D/K$ queue with service time T, can be calculated using the following expression due to Whitt [19] has suggested as follow

$$V[\tau] = (1-\rho^2)V[a] + (\frac{T}{K})^2(1-K^{-0.5}) \tag{35}$$

Note that (35) agrees with (30) when $K = 1$ and the arrival process is Poisson. However, Whitt's refinement does not seem to work well when the squared coefficient of variation of the arrival process, c_a^2 is large. In view of this, we will use the following modification :

$$V[\tau] = (1-\rho^2)V[a] + \frac{1+c_a^2}{2}(\frac{T}{K})^2(1-K^{-0.5}) \tag{36}$$

We now observe that the single node shown in Figure 10 is equivalent to the queueing system shown in case 3 of Figure 8. Following similar arguments as in the lower bound case, we can show that the variance of the interdeparture time of the queueing network in case 3 is an upper bound of the variance of the interdeparture time of the queueing network shown in case 2. In particular, from Theorem 2 we have that the i^{th} customer's waiting time in case 3 is smaller than that in case 2. Since the waiting time for each customer in case 3 decreases, the length of the busy period in case 3 is less than that in case 2. In view of this, the interdeparture times from the queueing network in case 2 are more regular than those in case 3. Therefore, the variance of the interdeparture time in case 3 is larger than the variance of the interdeparture time in the queueing network under study.

To summarize, case 1 of Figure 8 gives a lower bound on the variance of the interdeparture time, and an upper bound on the mean waiting time. Case 3 gives an upper bound on the variance of the interdeparture time, and a lower bound on the mean waiting time.

We note that the upper bound of the variance of the interdeparture time tends to the variance of the interarrival time as the number of tokens K increases. This is because, as K increases, the number of servers in the queueing system in Figure 9 increases as well and the waiting time tends to zero. Also, the lower

bound of the variance of the interdeparture time becomes equal to the variance of the interdeparture time of a $G/D/1$ queue with a service time s^*, when $\frac{Ks^*}{T} = 1$. This is because, in this case we have that $K = \frac{T}{s^*}$, and therefore, the service time $\frac{T}{K}$ of the $G/D/1$ queue that gives the lower bound becomes equal to s^*. Finally, let us consider the original two-node queueing network under study. As the number of tokens K increases, the queueing network behaves like a $G/D/1$ queue with service time s^*. Therefore, the variance of the interdeparture time tends to the same value as the lower bound, as $\frac{Ks^*}{T}$ goes to 1.

The behavior of these bounds in relation to the exact variance, as a function of $\frac{Ks^*}{T}$, is shown in Figure 11. Empirically, we have observed that the two bounds and the exact variance are concave functions of K, and that the exact variance tends to the upper bound as $\frac{Ks^*}{T}$ becomes small, and it tends to the lower bound as $\frac{Ks^*}{T}$ goes to 1.

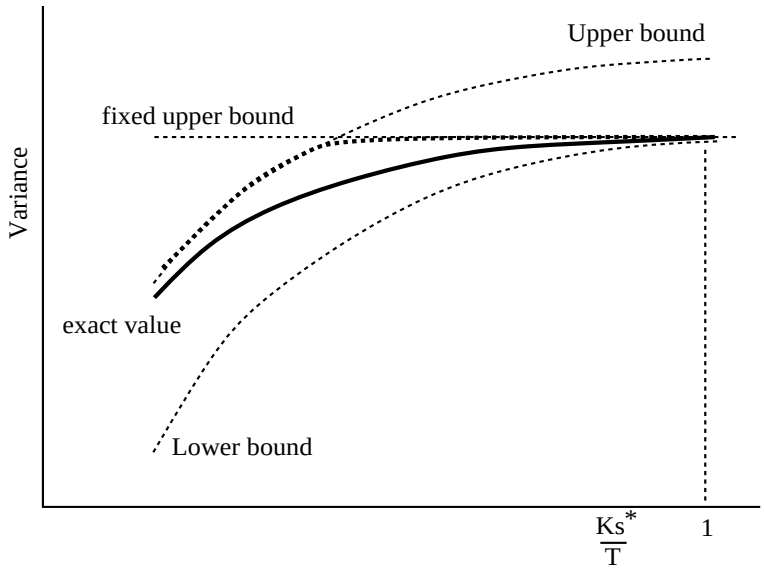

Figure 11: Bounds of variance

An alternative upper bound of the variance of the interdeparture time can be obtained using a $G/D/1$ queue with service time s^*. We refer to this upper bound as the *fixed upper bound*. Therefore, a better upper bound can be constructed

by taking the minimum of the variance of the interdeparture time of a $G/D/1$ queue with service time s^*, and the variance of the interdeparture time from the queueing system shown in Figure 10. This upper bound is indicated by a thick dotted line in Figure 11.

4.3 An Approximation to the Variance of the Interdeparture Time

In this subsection, we present an approximation to the variance of the interdeparture time of the semaphore controlled queueing network. Comparing Figure 11 and 3, we observe that the variance of the interdeparture time behaves in the opposite way than the mean waiting time. That is, it tends to the upper bound as $\frac{Ks^*}{T}$ becomes small, and it tends to the lower bound as $\frac{Ks^*}{T}$ goes to 1. Therefore, an approximation to the variance can be obtained by combining the upper and lower bounds using the same weight function $f(c_a^2, s^*, T, K)$ introduced in section 3.4. Let us define V_l and V_u to be the lower and upper bound of the variance respectively. Then,

$$V_a = V_u - f(c_a^2, s^*, T, K)(V_u - V_l) \tag{37}$$

The accuracy of the approximation was tested by comparing it against simulation. Below we report on three representative examples assuming Poisson arrivals and phase-type arrivals. For each example, we give the variance obtained by simulation, the approximation to the variance calculated by simulating the upper and lower bounds (simulated approximation), and the approximation to the variance obtained using expression (30) and (36) for calculating the upper and lower bounds (analytical approximation). As it can be seen, the approximation has a good accuracy.

Example 1 : Poisson arrivals with $\lambda = \frac{1}{7}$, $s^* = 2.5$ and $\overline{s} = 28$.

number of tokens	5	6	7	8	9	10
simulated approximation	30.02	37.17	40.45	41.89	42.50	42.63
simulated variance	29.23	36.54	40.09	41.68	42.38	42.63
analytical approximation	28.90	34.75	37.93	39.82	41.01	41.78

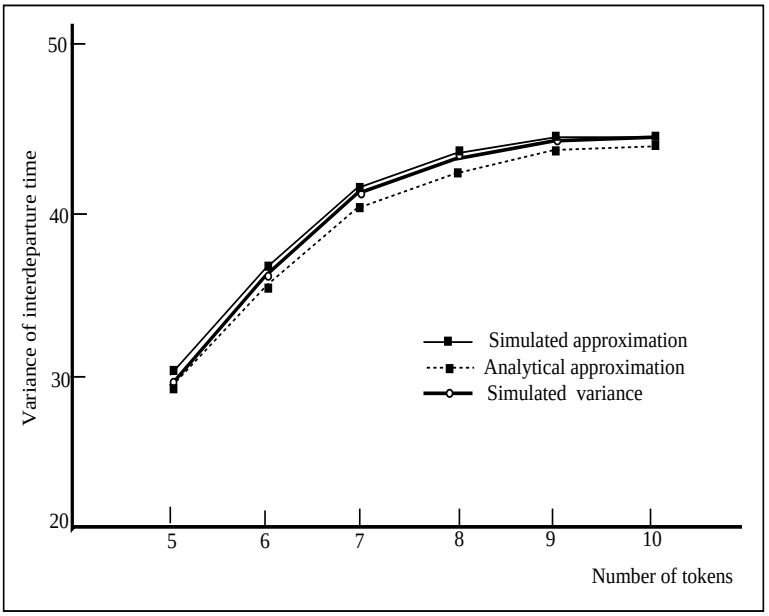

Figure 12: An approximation on the variance for Poisson($\frac{1}{7}$,2.5,28)

Example 2 : Erlang 2 phases with a phase arrival rate = $\frac{1}{3.5}$, $s^* = 2.5$ and $\overline{s} = 28$.

number of tokens	5	6	7	8	9	10
simulated approximation	17.66	21.05	22.49	22.83	22.84	22.85
simulated variance	17.21	20.81	22.15	22.56	22.67	22.69
analytical approximation	20.31	21.86	22.51	22.79	22.86	22.87

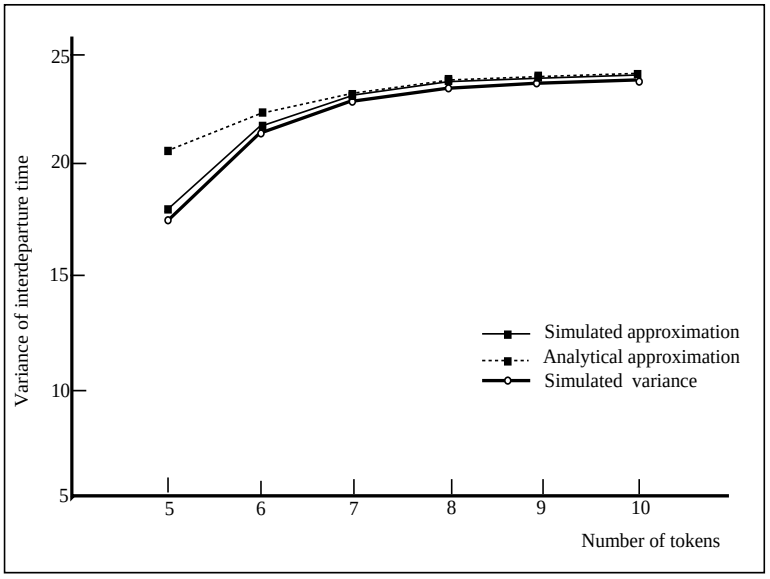

Figure 13: An approximation on the variance for Erlang(2,$\frac{1}{3.5}$,2.5,28)

Example 3 : hyper-exponential arrivals 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
simulated approximation	52.64	72.68	84.77	91.36	95.29	97.28
simulated variance	53.63	74.45	86.03	92.56	96.14	97.97
analytical approximation	51.78	69.64	80.06	86.66	91.14	94.30

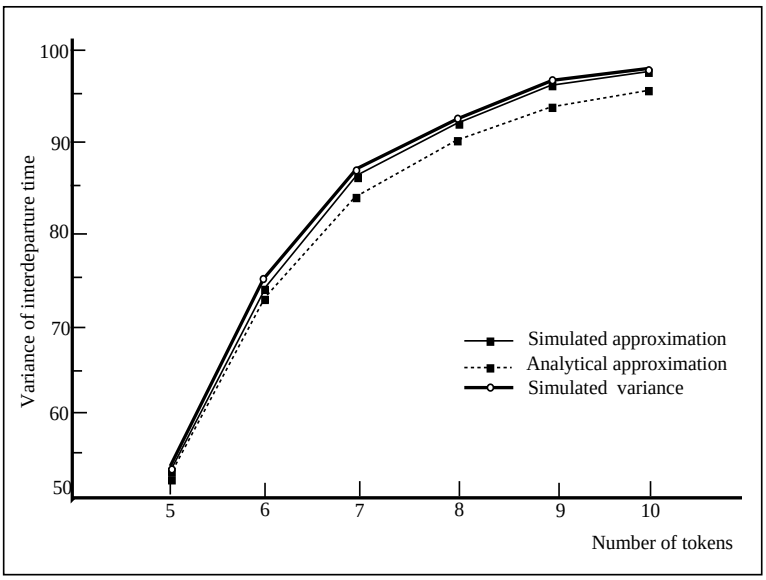

Figure 14: An approximation on the variance for Hyper($\frac{1}{3}$,$\frac{2}{3}$,$\frac{1}{15}$,$\frac{1}{3}$,2.5,28,2.31)

5 Conclusions

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

It was shown in Rhee and Perros [16] that the queueing network can be transformed into a simple queueing network involving only two nodes. Using this simple queueing network, upper and lower bounds on the mean waiting time were obtained.

In this paper, an approximation to the mean waiting time was suggested by weighing these two bounds appropriately. The departure process from the queueing network was also obtained. In particular, upper and lower bounds on the variance of the interdeparture time were obtained. An approximation to the variance was then calculated by combining these two bounds. Validations against simulation data showed that the approximations for the mean waiting time and the variance of the interdeparture time have a good accuracy.

References

[1] T. Altiok and Z. Kao, "Bounds for throughput 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 station with arbitrary input and regular service times", *Management Science*, Vol. 11, pp. 565-571, 1965.

[3] F. Baccelli and Z. Liu, "Comparison Properties of Stochastic Decision Free Petri Nets", *IEEE Trans. Auto. Control*, Vol. 37-12, pp.1905-1920, 1992.

[4] Y. Dallery, "Approximate analysis of general open queueing networks with restricted capacity", *Performance Evaluation*, Vol. 11, pp. 209-222, 1990.

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

[6] H.D. Friedman, "Reduction methods for Tandem queueing system." *Operations Research*, Vol. 13, 1965.

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

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

[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, 1976.

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

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

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

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

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

[15] 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.

[16] Y. Rhee and H.G. Perros,"Analysis of an Open Tandem Queueing Network with Population Constraint and Constant Service Times ",*European Journal of Operations Research*, to appear.

[17] M. Shalmon and M. Kaplan, "A tandem network of queues with deterministic service and intermediate arrivals", *Operations Research*, Vol. 32, No. 4, pp. 753-773, 1984.

[18] G.W. Shapiro and H.G. Perros, "Nested sliding window protocols with packet fragmentation", *IEEE Trans. Comm.*, Vol. 41, pp. 99-109, 1993.

[19] W. Witt, "Departure from a queue with many busy servers", *Mathematics of Operations Research*, Vol. 9, pp. 534-544, 1984.

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