BOUNDS AND HEURISTICS FOR ASSEMBLY-LIKE QUEUES

Wallace J. HOPP and John T. SIMON

Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, U.S.A.

Received 3 June 1988; revised 7 December 1988

Abstract

We consider an assembly system with exponential service times, and derive bounds for its average throughput and inventories. We also present an easily computed approximation for the throughput, and compare it to an existing approximation.

Keywords: Assembly-like queues, bounds, approximations.

1. Introduction

Assembly-like queues arise in many practical situations, including assembly lines in production plants (e.g. automobiles), mixing operations in chemical industries and data flow through computer systems (Dennis [6]). Despite their applicability, the literature on assembly-like queues is scarce, largely due to their analytical intractability.

In this paper we consider assembly-like queues with random service times. Such systems have been studied in the literature (Lipper and Sengupta [14]) and their randomness arises due to variability in processing times, especially in those processes in which randomness is inherent – for example, balancing of automobile shafts (Monden [15]). Randomness would also be a natural assumption in the case of dataflow models of computer systems mentioned above.

In a predominantly theoretical study, Harrison [9] considered an assembly-like queue whose input processes are independent renewal processes and with no restriction on the queue size of customers of each type. Under these assumptions, Harrison showed that the waiting time process does not converge in distribution to a non-defective limit. Latouche [13] showed that an assembly system with two Poisson arrivals and exponential service times, where the arrival rates depend on the excess of customers of one type over the other in such a way that the excess is bounded, is stable. Further, he indicated a matrix geometric technique based on the work of Neuts [16] for computing the stationary probability vector. Bonomi

[4] treated a similar system with more than two inputs, and gave an approximate procedure for computing throughput and mean queue lengths.

Bhat [3] analyzed finite capacity assembly-like queues, with emphasis on deriving the response time distributions assuming that the steady state probabilities are available. He did not address the computational aspects of obtaining the steady state probabilities. Lipper and Sengupta [14] considered a model which is essentially that studied here, and gave an approximate method for computing the throughput and mean inventory. In this model, each input process is Poisson with finite waiting space, and service times are exponential. This is a more realistic model of assembly systems than the "bounded excess" model of Latouche.

Although this model of assembly systems is clearly a Markov process, it generally requires a large state space and the 'curse of dimensionality' prevents us from obtaining analytical solutions in the case of reasonably large buffer sizes. In the absence of exact solutions, approximate methods and analytical bounds are the other alternatives for computing performance measures. The approximate method of Lipper and Sengupta provides one approach. However, because it is algorithmic in nature, their approach is not simple. It also does not provide error bounds. In this paper, we present some analytical bounds for throughput and inventory and also present an approximate solution very different from that of Lipper and Sengupta. Our approximate method is much easier to implement, and in some instances works better when compared to Lipper and Sengupta's for throughput. But our method is restricted to systems with two input sources, while Lipper and Sengupta's method works for more general systems. Their method also gives superior results for inventory.

2. Model description

The basic model of an assembly-like queueing system is depicted in fig. 1. Machines IM_1 and IM_2 are the input machines and AM is the assembly machine. We model the finite buffer space through bins. Machines IM_1 and IM_2 work to fill the bins, which then travel to AM where they are emptied. The empty bins travel back to machines IM_1 and IM_2 respectively. Travel times are considered to be negligible.

For machine IM_1 to function it must have at least one empty bin in front of it. Machine IM_2 operates likewise. The bins of the two types do not mix – a full bin that came from machine IM_1 returns to machine IM_1 when it is emptied. For machine AM to function, there must be at least one full bin in buffer B_1 (i.e. from machine IM_1) and at least one full bin in buffer B_2. Thus, this model also depicts a "pull" or "kanban" inventory control system.

Each bin may carry one or many components. For the assembly operation,we may need two components of one kind and one of the other, and here we assume

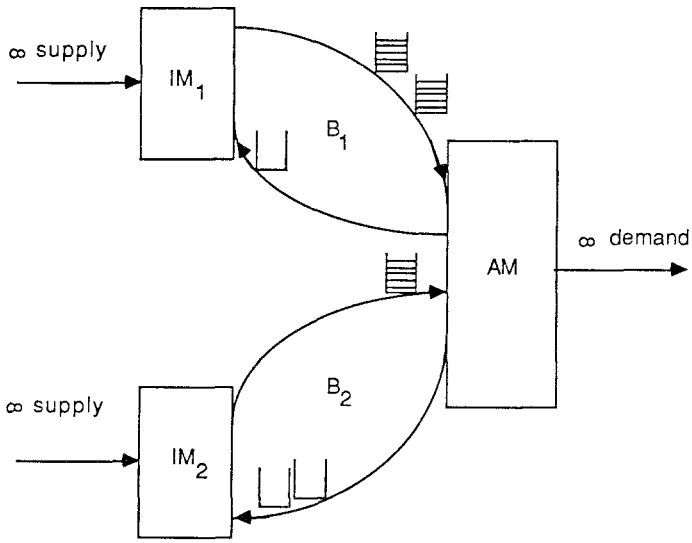

Fig. 1.

that the size of the bin is suitably scaled that exactly one bin of each type of components is used for assembly.

We have made the implicit assumption that a full bin remains at machine AM until the machine AM completes its operation on the contents of that bin. Alternatively we could assume that a bin is released as soon as machine AM starts operation on its contents (i.e., the contents of the bin is transferred to the machine and the bin is freed). But this can be shown to be equivalent to the previous model with one additional bin in each buffer. Hence, there is no need to analyze this model separately.

For the purpose of analysis, we now make the assumption that the service times are independent exponential random variables, with rates λ_1, λ_2 and μ for machines IM_1, IM_2 and AM, respectively. Let $N_1(t)$ be the number of bins in buffer B_1 waiting for service from machine AM at time t and let $N_2(t)$ be those in B_2 waiting for machine AM. Define a state (n_1, n_2) to mean that $N_1(t) = n_1$ and $N_2(t) = n_2$. Let the total number of bins in buffer B_1 be K_1 and that in B_2 be K_2. These are referred to as the buffer sizes or capacities.

The parameters λ_1, λ_2, μ, K_1 and K_2 completely describe this model. The performance measures we are concerned with are the steady state mean throughput θ, and the steady state average inventory in each buffer, denoted $\mathscr{I}_1$ and $\mathscr{I}_2$. θ is defined to be the mean number of service completions of machine AM in unit time in steady state (actually, the mean steady state throughput of machines IM_1, IM_2 and AM are all equal). $\mathscr{I}_1$ is defined to be the steady state mean queue length of bins in buffer B_1 waiting for service at machine AM. $\mathscr{I}_2$ is similarly defined. In all the discussion to follow, we assume steady state operating

conditions. So *mean throughput* stands for the *steady state mean throughput* and likewise for *mean inventories*.

To facilitate our discussion we make use of the following notation: $\{\lambda/\mu/1/K\}$ stands for an M/M/1/K queue with inter-arrival and service time rates given by λ and μ respectively. $\theta\{\lambda/\mu/1/K\}$, $\mathcal{I}\{\lambda/\mu/1/K\}$ and $p_0\{\lambda/\mu/1/K\}$ stand for the mean throughput, mean queue length and empty probability of $\{\lambda/\mu/1/K\}$ respectively. $\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ stands for the assembly system described above, and $\theta\{\lambda_1/\lambda_2/\mu/IK_1/K_2\}$, $\mathcal{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ and $\mathcal{I}_2\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ stand for the mean throughput and the mean inventories of $\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ respectively (where there is no ambiguity, these are sometimes abbreviated as θ, $\mathcal{I}_1$ and $\mathcal{I}_2$).

Letting $\rho = \lambda/\mu$, from Gross and Harris [8] we have

$$\theta\{\lambda/\mu/1/K\} = \begin{cases} \left(1-(1-\rho)/(1-\rho^{K+1})\right)\mu & \text{if } \rho \neq 1 \\ (K/(K+1))\mu & \text{if } \rho = 1 \end{cases}$$

$$\mathcal{I}\{\lambda/\mu/1/K\} = \begin{cases} \rho\left[1-(K+1)\rho^K + K\rho^{K+1}\right] \\ \quad /\left[(1-\rho)(1-\rho^{K+1})\right] & \text{if } \rho \neq 1 \\ K/2 & \text{if } \rho = 1 \end{cases}$$

$$p_0\{\lambda/\mu/1/K\} = \begin{cases} (1-\rho)/(1-\rho^{K+1}) & \text{if } \rho \neq 1 \\ 1/(K+1) & \text{if } \rho = 1. \end{cases}$$

3. Equivalence of the assembly system to a transfer line

The first result we present is that the assembly system depicted in fig. 1 is equivalent (the nature of the equivalence is stated in theorem 1) to a transfer line of tandem queues with blocking. This equivalence is of practical interest because considerable effort has been devoted to the analysis of tandem queues (see Altiok [1], Buzacott [5], Gershwin and Schick [7], Hatcher [10], Hillier and Boling [11] and Hunt [12]).

Consider the three machine transfer line with finite buffers between the machines shown in fig. 2. Machine IM_1' works as long as there is an empty bin in buffer B_1'. For machine AM′ to function, there must be a full bin in buffer B_1'

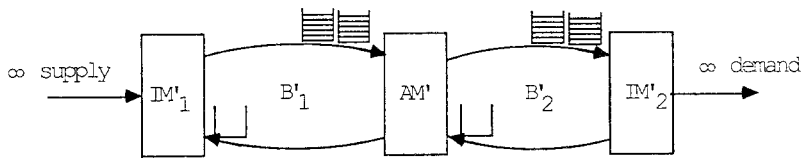

Fig. 2.

and an empty bin in buffer B_2'. Machine IM_2' works as long as there is a full bin in B_2'. We define the number of bins in B_1' and B_2' to be K_1 and K_2, respectively, the service times at IM_1', IM_2' and AM′ to be independent exponential random variables with rates λ_1, λ_2 and μ, respectively, $N_1'(t)$ to be the number of full bins in buffer B_1' and $N_2'(t)$ to be the number of *empty* bins in buffer B_2' at time t. Clearly this represents an ordinary three machine transfer line with two finite buffers in between the machines.

THEOREM 1

The process $\{N_1'(t), N_2'(t);\ t > 0\}$ is stochastically equivalent to the process $\{N_1(t), N_2(t);\ t \geqslant 0\}$ described above in section 3.

Proof

Their equivalence can be seen by starting both the processes with the same initial state, and using the same sample path in both processes. The fact that the machines have exponential service times is not used here. As long as the successive service times at machines IM_1 and IM_1' are the same, IM_2 and IM_2' are the same, and AM and AM′ are the same, this equivalence holds. □

As a by-product, we see that the throughputs of both the assembly system and the transfer line are the same. Also the average steady state inventory in B_1' is given by $\mathscr{I}_1$, and that in B_2' is given by $K_2 - \mathscr{I}_2$.

A version of this equivalence, where the processing times are deterministic but machines are subject to failure, is given in Ammar [2].

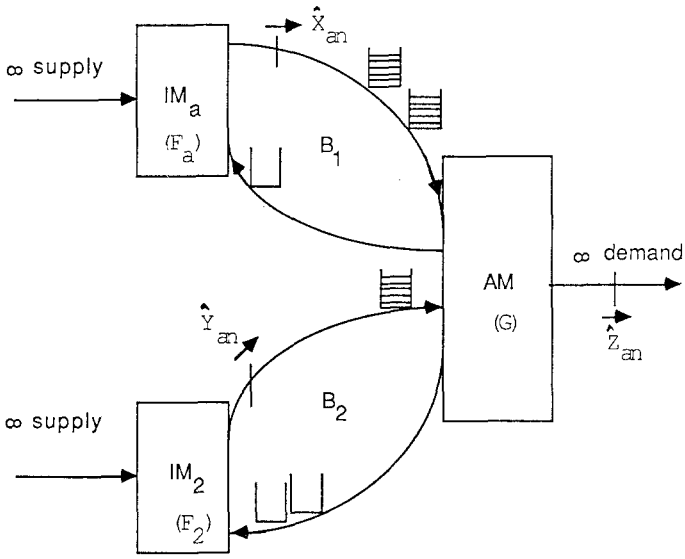

Fig. 3.

4. Upper bound for throughput

Let $\{F_1/F_2/G/K_1/K_2\}$ denote an assembly system shown in fig. 3, where the successive service times of IM_1 are independent and identically distributed random variables (iid rv's) with cumulative distribution function (cdf) F_1, service times of IM_2 are iid rv's with cdf F_2, service times of AM are iid rv's with cdf G, buffer B_1 has a capacity of K_1 bins and B_2 has K_2 bins. F_1, F_2 and G need not be exponential distributions. Let $\theta\{F_1/F_2/G/K_1/K_2\}$ denote the steady state average throughput of $\{F_1/F_2/G/K_1/K_2\}$. In the following, we use $\leqslant^{st}$ to mean "stochastically less than", as defined in Ross [17].

LEMMA 1

$F_a \leqslant^{st} F_b \Rightarrow \theta\{F_a/F_2/G/K_1/K_2\} \geqslant \theta\{F_b/F_2/G/K_1/K_2\}$.

Proof

We generate the successive service times at IM_a, IM_b, IM_2 and AM as follows:

$$\hat{S}_{a1}, \hat{S}_{a2}, \hat{S}_{a3}, \hat{S}_{a4}, \ldots \sim F_a$$

$$\hat{S}_{b1}, \hat{S}_{b2}, \hat{S}_{b3}, \hat{S}_{b4}, \ldots \sim F_b$$

$$\hat{S}_{21}, \hat{S}_{22}, \hat{S}_{23}, \hat{S}_{24}, \ldots \sim F_2$$

$$\hat{T}_1, \hat{T}_2, \hat{T}_3, \hat{T}_4, \ldots \sim G$$

where $\hat{S}_{an}$ and $\hat{S}_{bn}$ are generated by choosing $\hat{U}_1, \hat{U}_2, \hat{U}_3, \hat{U}_4, \ldots$ to be independent random variables uniformly distributed in [0, 1], and taking $\hat{S}_{an} = F_a^{-1}(\hat{U}_n)$ and $\hat{S}_{bn} = F_b^{-1}(\hat{U}_n)$ (see fig. 4). Clearly $\hat{S}_{an} \sim F_a$ and $\hat{S}_{bn} \sim F_b$. Furthermore, since $F_a \leqslant^{st} F_b$, it follows that

$$\hat{S}_{an} \leqslant \hat{S}_{bn} \text{ for all } n. \tag{1}$$

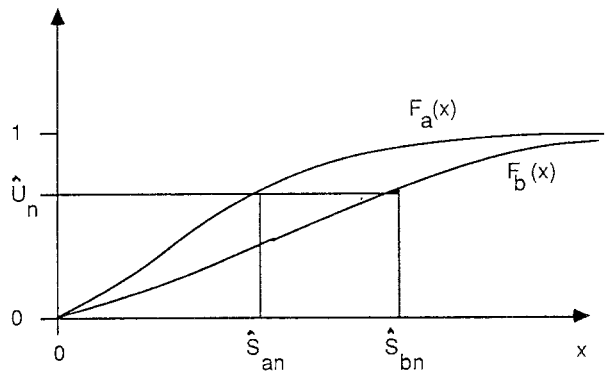

Fig. 4.

Let the successive service completion times of IM_a, IM_2 and AM of the system $\{F_a/F_2/G/K_1/K_2\}$ be denoted by (see fig. 3)

$$\hat{X}_{a1}, \hat{X}_{a2}, \hat{X}_{a3}, \hat{X}_{a4}, \ldots$$
$$\hat{Y}_{a1}, \hat{Y}_{a2}, \hat{Y}_{a3}, \hat{Y}_{a4}, \ldots$$
$$\hat{Z}_{a1}, \hat{Z}_{a2}, \hat{Z}_{a3}, \hat{Z}_{a4}, \ldots$$

respectively, and those of IM_b, IM_2 and AM of the system $\{F_b/F_2/G/K_1/K_2\}$ be denoted by

$$\hat{X}_{b1}, \hat{X}_{b2}, \hat{X}_{b3}, \hat{X}_{b4}, \ldots$$
$$\hat{Y}_{b1}, \hat{Y}_{b2}, \hat{Y}_{b3}, \hat{Y}_{b4}, \ldots$$
$$\hat{Z}_{b1}, \hat{Z}_{b2}, \hat{Z}_{b3}, \hat{Z}_{b4}, \ldots$$

respectively. Assuming that both systems start with all empty bins, we have

$$\hat{X}_{a1} = \hat{S}_{a1} \tag{2}$$
$$\hat{X}_{a\,n+1} = \hat{X}_{an} + \hat{S}_{a\,n+1} \quad \text{for } 1 \leqslant n \leqslant K_1 - 1 \tag{2'}$$
$$\hat{X}_{a\,n+1} = \max\{\hat{X}_{an}, \hat{Z}_{a\,n+1-K_1}\} + \hat{S}_{a\,n+1} \quad \text{for } n \geqslant K_1 \tag{2''}$$
$$\hat{Y}_{a1} = \hat{S}_{21} \tag{3}$$
$$\hat{Y}_{a\,n+1} = \hat{Y}_{an} + \hat{S}_{2\,n+1} \quad \text{for } 1 \leqslant n \leqslant K_2 - 1 \tag{3'}$$
$$\hat{Y}_{a\,n+1} = \max\{\hat{Y}_{an}, \hat{Z}_{a\,n+1-K_2}\} + \hat{S}_{2\,n+1} \quad \text{for } n \geqslant K_2 \tag{3''}$$
$$\hat{Z}_{a\,n+1} = \max\{\hat{Z}_{an}, \hat{X}_{a\,n+1}, \hat{Y}_{a\,n+1}\} + \hat{T}_{n+1} \quad \text{for } n \geqslant 1 \tag{4}$$

for $\{F_a/F_2/G/K_1/K_2\}$, and

$$\hat{X}_{b1} = \hat{S}_{b1} \tag{5}$$
$$\hat{X}_{b\,n+1} = \hat{X}_{bn} + \hat{S}_{b\,n+1} \quad \text{for } 1 \leqslant n \leqslant K_1 - 1 \tag{5'}$$
$$\hat{X}_{b\,n+1} = \max\{\hat{X}_{bn}, \hat{Z}_{b\,n+1-K_1}\} + \hat{S}_{b\,n+1} \quad \text{for } n \geqslant K_1 \tag{5''}$$
$$\hat{Y}_{b1} = \hat{S}_{21} \tag{6}$$
$$\hat{Y}_{b\,n+1} = \hat{Y}_{bn} + \hat{S}_{2\,n+1} \quad \text{for } 1 \leqslant n \leqslant K_2 - 1 \tag{6'}$$
$$\hat{Y}_{b\,n+1} = \max\{\hat{Y}_{bn}, \hat{Z}_{b\,n+1-K_2}\} + \hat{S}_{2\,n+1} \quad \text{for } n \geqslant K_2 \tag{6''}$$
$$\hat{Z}_{b\,n+1} = \max\{\hat{Z}_{bn}, \hat{X}_{b\,n+1}, \hat{Y}_{b\,n+1}\} + \hat{T}_{n+1} \quad \text{for } n \geqslant 1 \tag{7}$$

for $\{F_a/F_2/G/K_1/K_2\}$.

Suppose that for some positive integer m, we have

$$\hat{X}_{am} \leqslant \hat{X}_{bm} \tag{8}$$
$$\hat{Y}_{am} \leqslant \hat{Y}_{bm} \tag{9}$$
$$\hat{Z}_{am} \leqslant \hat{Z}_{bm}. \tag{10}$$

Clearly it holds for $m = 1$. By eqs. (2), (2′), (2″), (5), (5′), (5″), (8) and (1), it follows that $\hat{X}_{a\,m+1} \leqslant \hat{X}_{b\,m+1}$. Likewise $\hat{Y}_{a\,m+1} \leqslant \hat{Y}_{b\,m+1}$. These two inequalities, together with eqs. (4), (7) and (10) yield $\hat{Z}_{a\,m+1} \leqslant \hat{Z}_{b\,m+1}$.

Thus we have shown by induction that for all $n \geqslant 1$, we have

$$\hat{Z}_{an} \leqslant \hat{Z}_{bn}.$$

Since $n/\hat{Z}_{an}$ converges to $\theta\{F_a/F_2/G/K_1/K_2\}$ with probability one, and likewise $n/\hat{Z}_{bn}$ converges to $\theta\{F_b/F_2/G/K_1/K_2\}$, the lemma is proved. □

COROLLARY 1

$$\lambda_a \leqslant \lambda_b \Rightarrow \theta\{\lambda_a/\lambda_2/\mu/K_1/K_2\} \leqslant \theta\{\lambda_b/\lambda_2/\mu/K_1/K_2\}$$

Proof

Let F_a be an exponential distribution with rate λ_a, and F_b an exponential distribution with rate λ_b, $\lambda_a \leqslant \lambda_b \Rightarrow F_b \leqslant^{\text{st}} F_a$, and hence the conclusion follows from lemma 1. □

COROLLARY 2

$$\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \leqslant \theta\{\lambda_1/\mu/1/K_1\}$$

Proof

$\{\lambda_1/\mu/1/K_1\}$ is the same as $\{\lambda_1/\infty/\mu/K_1/K_2\}$. If F_1 is an exponential distribution with rate λ_1, and G an exponential distribution with rate μ, the latter may also be written as $\{F_1/I/G/K_1/K_2\}$ where I is the unit step function at zero. $F \leqslant^{\text{st}} I$ for any cdf F of a positive random variable, $\theta\{F_1/F_2/G/K_1/K_2\} = \theta\{F_2/F_1/G/K_2/K_1\}$ by symmetry, and hence the conclusion follows from lemma 1. □

LEMMA 2

As μ increases to ∞, $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ increases monotonically to $\theta\{\lambda_1/\lambda_2/1/K_1 + K_2\}$.

Proof

Proof is analogous to the proof of lemma 1 and corollary 2. □

From corollary 2 and lemma 2, we have that $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ is always bounded as follows:

$$\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \leqslant \theta\{\lambda_1/\mu/1/K_1\}$$
$$\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \leqslant \theta\{\lambda_2/\mu/1/K_2\}$$
$$\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \leqslant \theta\{\lambda_1/\lambda_2/1/K_1 + K_2\}.$$

Therefore we have the following theorem.

THEOREM 2

$\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \leqslant \theta_{ub}$, where

$$\theta_{ub} = \min\{\theta\{\lambda_1/\mu/1/K_1\}, \theta\{\lambda_2/\mu/1/K_2\}, \theta\{\lambda_1/\lambda_2/1/K_1 + K_2\}\}. \quad \square$$

Also from corollary 2 and lemma 2, we see that asymptotically, as $\lambda_1 \to \infty$ or as $\lambda_2 \to \infty$ or as $\mu \to \infty$, the upper bound becomes tight. So if we have any one of λ_1, λ_2 or μ large compared to the others, this upper bound will be fairly close to the actual throughput.

Notice that even if the machines in the assembly system had general service times instead of exponentially distributed service times, we could derive an upper bound to the throughput analogous to that in theorem 2. However, for computing the upper bound, we would need the throughput of a GI/G/1/K queue.

5. Lower bound for throughput

We now derive two different lower bounds on throughput. First we need the following result. Recall that $N_1(t)$ represents the number of full bins in buffer B_1 of the assembly system. Let N_1 be the number of full bins in B_1 of the assembly system in the steady state.

LEMMA 3

$P(N_1 = 0) \leqslant p_0\{\lambda_1/\mu/1/K_1\}$

Proof

Define $T_1(t)$ to be the time during $[0, t]$ when $N_1(t)$ is equal to zero. Let $T_0(t)$ be the time during $[0, t]$ that the queue $\{\lambda_1/\mu/1/K_1\}$ is empty. By an argument similar to the one in the proof of lemma 1, we can show that $\{T_1(t)\}$ is stochastically less than $\{T_0(t)\}$. Since

$$P(N_1 = 0) = \lim_{t\to\infty} \frac{T_1(t)}{t}$$

and

$$p_0\{\lambda_1/\mu/1/K_1\} = \lim_{t\to\infty} \frac{T_0(t)}{t},$$

the lemma follows. $\square$

This lemma leads directly to our first lower bound.

LEMMA 4

$\theta_{lb1} \equiv \mu[1 - p_0\{\lambda_1/\mu/1/K_1\} - p_0\{\lambda_2/\mu/1/K_2\}] \leqslant \theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$

Proof

$$\begin{aligned}\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\} &= \mu(1 - P(N_1 = 0 \text{ or } N_2 = 0)) \\ &\geqslant \mu(1 - P(N_1 = 0) - P(N_2 = 0)) \\ &\geqslant \mu(1 - p_0\{\lambda_1/\mu/1/K_1\} - p_0\{\lambda_2/\mu/1/K_2\})\end{aligned}$$

by the previous lemma. □

Notice that as $\lambda_1 \to \infty$, θ_{lb1} increases monotonically to $\theta\{\lambda_1/\mu/1/K_2\}$ and as $\lambda_2 \to \infty$, θ_{lb1} increases monotonically to $\theta\{\lambda_1/\mu/1/K_1\}$, so the bound is tight for large λ_1 or λ_2. However, as $\mu \to \infty$, $\theta_{lb1} \to -\infty$, which implies that this bound will perform poorly for the case where $\mu \gg \lambda_1, \lambda_2$. However, if the assembly operation is the bottleneck, we will have $\mu \leqslant \lambda_1, \lambda_2$. Hence, this bound may be useful in practice.

Next, we derive another lower bound on throughput by considering an assembly system in which each machine processes k bins and shuts off until the other machines have also completed k bins, where

$$k = \left\lfloor \frac{\min\{K_1, K_2\}}{2} \right\rfloor.$$

($\lfloor x \rfloor$ is defined to be the largest integer less than or equal to x.) After each machine completes k bins, the process is repeated. If we start with k full bins in each buffer in front of machine AM, $K_1 - k$ bins in front of machine IM_1, $K_2 - k$ bins in front of machine IM_2, then processing k bins at each machine returns the system to this same state. The throughput of this system is a lower bound on $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ and is given by

$$\theta_{lb2} = \frac{k}{E[\max\{\text{Erlang}(\lambda_1, k), \text{Erlang}(\lambda_1, k), \text{Erlang}(\mu, k)\}]},$$

where Erlang(λ, k) is a random variable which is the sum of k independent exponential random variables each with rate λ.

LEMMA 5

$\theta_{lb2} \leqslant \theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$

Proof

The throughput of the new system is easily found using the renewal reward process of Ross [17] to be θ_{lb2}. That this throughput is a lower bound to the throughput of the original assembly system is shown by an argument analogous to the proof of lemma 1. □

θ_{lb2} can be computed iteratively, as outlined in appendix A. (In the case of an assembly system with general service times instead of exponentially distributed

service times, we could derive a similar lower bound for the throughput. However, instead of the iterative computation given in appendix A, we would have to compute the appropriate convolutions of distributions.) We have thus proven the following theorem.

THEOREM 3

$\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geqslant \theta_{lb}$, where $\theta_{lb} = \max\{\theta_{lb1}, \theta_{lb2}\}$. □

HEURISTIC LOWER BOUND FOR THROUGHPUT

We now give an approximate method for computing throughput that, while not a demonstrable lower bound, virtually always underestimates throughput. We will make use of this "heuristic lower bound" to give a simple approximation of throughput.

To motivate the heuristic lower bound, consider two separate transfer lines as shown in fig. 5. Let $N_1(t)$ be the number of full bins in buffer B_1 at time t and $N_2(t)$ that in B_2. The machines IM_1, IM_2, AM_a and AM_b are exponential servers with rates λ_1, λ_2, μ and μ respectively. The buffer size of B_1 is K_1 and that of B_2 is K_2.

It is clear that if machine AM_b is deactivated whenever $N_1(t) = 0$ and machine AM_a is deactivated whenever $N_2(t) = 0$, and the sample path of successive service times for AM_a and AM_b are the same, we again have $\{N_1(t), N_2(t);\ t \geqslant 0\}$ to be the same Markov process as we had earlier in the assembly system. Suppose, however, that transfer line b operates without any influence from the transfer line a, but machine AM_a is deactivated whenever $N_2(t) = 0$. We let θ_a

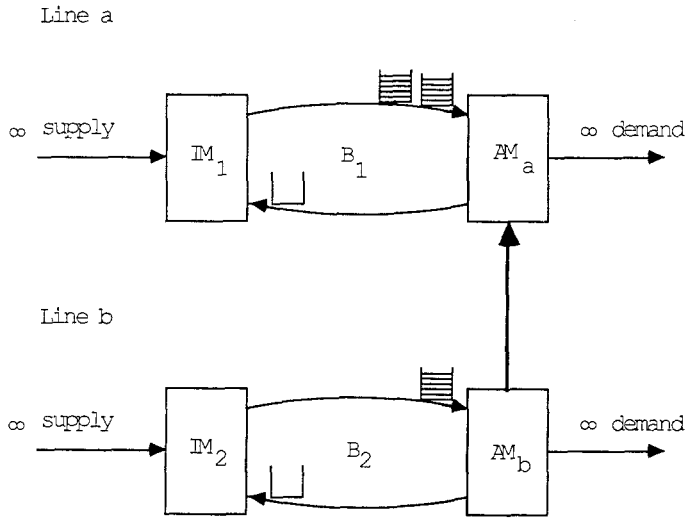

Fig. 5.

represent the throughput of line a under these conditions and θ_b represent the throughput of line b when line a operates independent of line b and AM_b is deactivated whenever $N_1(t) = 0$. We can demonstrate that these throughputs represent lower bounds on the actual throughput.

LEMMA 6

$$\theta_a \leqslant \theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$$
$$\theta_b \leqslant \theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$$

Proof

The proof follows from a similar argument to that given in the proof of lemma 1. □

Unfortunately, computing θ_a and θ_b is essentially as difficult as computing $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$, so these bounds are not of practical use. To develop an easily computable approximation to these bounds, further suppose that the effect of slowing down of line a due to $N_2(t)$ being zero in line b is captured by reducing the service rate of AM_1 by a factor of $P(N_2 = 0) = p_0\{\lambda_2/\mu/1/K_2\}$ (in the steady state). To the extent that this is true, an approximation to θ_a is given by $\theta\{\lambda_1/\mu[1 - p_0\{\lambda_2/\mu/1/K_2\}]/1/K_1\}$. Hence, this approximation should serve as a lower bound on the actual throughput, $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$. Extensive computations, summarized in section 8 show that it does indeed consistently underestimate the throughput. Additionally, this approximation is very simple to compute. We simply compute $p_0\{\lambda_2/\mu/1/K_2\}$ (using standard results for the M/M/1/K queue), set $\mu' = \mu[1 - p_0\{\lambda_2/\mu/1/K_2\}]$, and then compute $\theta\{\lambda_1/\mu'/1/K_1\}$ (again using the standard M/M/1/K queue results).

We define a heuristic lower bound for the throughput to be the larger of the approximations θ_a and θ_b:

$$\theta_{\text{hlb}} = \max\{\theta\{\lambda_1/\mu[1 - p_0\{\lambda_2/\mu/1/K_2\}]/1/K_1\},$$
$$\theta\{\lambda_2/\mu[1 - p_0\{\lambda_1/\mu/1/K_1\}]/1/K_2\}\}.$$

As $\lambda_1 \to \infty$, $\theta_{\text{hlb}} \uparrow \theta\{\lambda_2/\mu/1/K_2\}$. Likewise, as $\lambda_2 \to \infty$, $\theta_{\text{hlb}} \uparrow \theta\{\lambda_1/\mu/1/K_1\}$. θ_{ub} and θ_{lb} also exhibited the same behavior, so for large values of λ_1 or λ_2, the bounds are all tight. But as $\mu \to \epsilon$, $\theta_{\text{hlb}} \uparrow \max\{\theta\{\lambda_1/\lambda_2/1/K_1\}, \theta\{\lambda_2/\lambda_1/1/K_2\}\}$. In this case, θ_{ub} would be closer to $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ than other bounds.

6. Upper bound on inventory

Any lower bound on the throughput yields an upper bound on the inventory. We state this as our next lemma. Clearly it is not an efficient upper bound.

LEMMA 7

For any lower bound θ_{lb} on $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$,

$$\mathcal{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \leqslant K_1 - \frac{\theta_{lb}}{\lambda_1}$$

$$\mathcal{I}_2\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \leqslant K_2 - \frac{\theta_{lb}}{\lambda_2}.$$

Proof

Recall that N_1 represents the number of full bins in buffer B_1 of the assembly system in the steady state. Throughput of the assembly system, being also the throughput of the input machine IM_1, can be written as

$$\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\} = \lambda_1[1 - P(N_1 = K_1)].$$

It follows that $P(N_1 \neq K_1) = 1 - P(N_1 = K_1) \geqslant \theta_{lb}/\lambda_1$. Now

$$\begin{aligned}\mathcal{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} &= \sum_{n=1}^{K_1} nP(N_1 = n)\\ &\leqslant K_1 P(N_1 = K_1) + (K_1 - 1)P(N_1 \neq K_1)\\ &\leqslant K_1 - \theta_{lb}/\lambda_1.\end{aligned}$$

The second inequality follows analogously to this one. □

7. Lower bound on inventory

Similar to the upper bound on the inventories, any upper bound on the throughput gives us a lower bound on inventory.

LEMMA 8

For any upper bound θ_{ub} on $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$,

$$\mathcal{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geqslant \left(1 - \frac{\theta_{ub}}{\lambda_1}\right)K_1 \equiv \mathcal{I}_{lb1a}$$

$$\mathcal{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geqslant \left(1 - \frac{\theta_{ub}}{\lambda_2}\right)K_2 \equiv \mathcal{I}_{lb2a}$$

Proof

As in lemma 7, $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\} = \lambda_1[1 - P(N_1 = K_1)]$. If $\theta_{ub} \geqslant \theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$, we have

$$P(N_1 = K_1) \geqslant 1 - \theta_{ub}/\lambda_1.$$

Now,

$$\begin{aligned}\mathscr{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} &= \sum_{n=1}^{K_1} nP(N_1 = n)\\ &\geqslant K_1 P(N_1 = K_1)\\ &\geqslant K_1(1 - \theta_{ub}/\lambda_1).\end{aligned}$$

The second inequality follows analogously. □

Our next bound compares the inventory of the assembly system to that of a corresponding transfer line.

LEMMA 9

$$\mathscr{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geqslant \mathscr{I}\{\lambda_1/\mu/1/K_1\} \equiv \mathscr{I}_{lb1b}$$
$$\mathscr{I}_2\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geqslant \mathscr{I}\{\lambda_2/\mu/1/K_2\} \equiv \mathscr{I}_{lb2b}$$

Proof

The proof of this lemma is analogous to the proof of lemma 1. □

LEMMA 10

Consider an assembly system with parameters λ_1, λ_2, K_1 and K_2, with $\mu = \infty$. Let the mean inventories for this system be $\mathscr{I}_{lb1c}$ and $\mathscr{I}_{lb2c}$. Then $\mathscr{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geqslant \mathscr{I}_{lb1c}$ and $\mathscr{I}_2\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geqslant \mathscr{I}_{lb2c}$.

Proof

Again this is proved using arguments similar to those used in the proof of lemma 1. □

Computation of $\mathscr{I}_{lb1c}$ and $\mathscr{I}_{lb2c}$ are given in appendix B.

The preceding three lemmas yield the following lower bound for the inventories in the assembly system.

THEOREM 4

$$\mathscr{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geqslant \max\{\mathscr{I}_{lb1a}, \mathscr{I}_{lb1b}, \mathscr{I}_{lb1c}\}$$
$$\mathscr{I}_2\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geqslant \max\{\mathscr{I}_{lb2a}, \mathscr{I}_{lb2b}, \mathscr{I}_{lb2c}\}. \qquad \square$$

Heuristic upper bound for inventory

Mimicking the heuristic for lower bound on throughput, we can derive the following heuristic for inventory in each buffer of the assembly system:

$$\mathscr{I}_{hub1} = \mathscr{I}\{\lambda_1/\mu[1 - p_0\{\lambda_2/\mu/1/K_2\}]/1/K_1\}$$
$$\mathscr{I}_{hub2} = \mathscr{I}\{\lambda_2/\mu[1 - p_0\{\lambda_1/\mu/1/K_1\}]/1/K_2\}.$$

Table 1
Computational results

K_1	K_2	λ_1	λ_2	μ	θ_{act}	θ_{Lip}	% err	θ_{ub}	θ_{hlb}	θ_{apr}	% err
4	4	2.0	2.0	1.0	0.947	0.943	−0.4	0.968	0.940	0.954	0.7
4	4	1.4	1.4	1.0	0.865	0.859	−0.7	0.909	0.845	0.877	1.4
4	4	1.0	1.0	1.0	0.736	0.731	−0.7	0.800	0.703	0.752	2.2
4	4	0.6	0.6	1.0	0.502	0.506	0.8	0.566	0.466	0.516	2.8
4	4	0.2	0.2	1.0	0.176	0.182	3.4	0.200	0.160	0.180	2.3
4	4	0.05	0.05	1.0	0.044	0.046	4.5	0.050	0.040	0.045	2.3
7	7	2.0	2.0	1.0	0.993	0.992	−0.1	0.996	0.992	0.994	0.1
7	7	1.4	1.4	1.0	0.987	0.986	−0.1	0.993	0.986	0.990	0.3
7	7	1.0	1.0	1.0	0.831	0.829	−0.2	0.875	0.810	0.843	1.4
7	7	0.6	0.6	1.0	0.549	0.560	2.0	0.593	0.522	0.558	1.6
7	7	0.2	0.2	1.0	0.186	0.194	4.3	0.200	0.175	0.188	1.1
7	7	0.05	0.05	1.0	0.047	0.049	4.2	0.050	0.044	0.047	0.0
10	10	2.0	2.0	1.0	0.999	0.999	0.0	1.000	0.999	1.000	0.1
10	10	1.4	1.4	1.0	0.983	0.981	−0.2	0.990	0.981	0.986	0.3
10	10	1.0	1.0	1.0	0.876	0.875	−0.1	0.909	0.860	0.885	1.0
10	10	0.6	0.6	1.0	0.566	0.581	2.7	0.599	0.545	0.572	1.1
10	10	0.2	0.2	1.0	0.190	0.198	4.2	0.200	0.182	0.191	0.5
10	10	0.05	0.05	1.0	0.048	0.050	4.2	0.050	0.045	0.048	0.0
2	2	1.0	1.0	1.0	0.578			0.667	0.526	0.597	3.3
2	2	0.2	1.8	1.0	0.194			0.194	0.192	0.193	−0.5
2	2	0.2	0.2	1.8	0.157			0.197	0.132	0.165	5.1
2	2	1.8	1.8	0.2	0.196			0.198	0.196	0.197	0.5
14	14	1.0	1.0	1.0	0.908			0.933	0.897	0.915	0.8
14	14	0.2	1.8	1.0	0.200			0.200	0.200	0.200	0.0
14	14	0.2	0.2	1.8	0.193			0.200	0.187	0.194	0.5
14	14	1.8	1.8	0.2	0.200			0.200	0.200	0.200	0.0
2	20	1.0	1.0	1.0	0.667			0.667	0.667	0.667	0.0
2	20	0.2	1.8	1.0	0.194			0.194	0.195	0.194	0.0
2	20	0.2	0.2	1.8	0.190			0.198	0.189	0.194	2.1
2	20	1.8	1.8	0.2	0.198			0.198	0.198	0.198	0.0

While this approximation tends to overestimate inventory, it does not do so consistently – there are cases where $\mathscr{I}_{\text{hub1}} \leqslant \mathscr{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ or $\mathscr{I}_{\text{hub2}} \leqslant \mathscr{I}_2\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$.

8. Computational results

From extensive computations, we found that the average of θ_{ub} and θ_{hlb} gives a good approximation to $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$. A sample of computational results for the approximation to the throughput is presented in table 1. In this table, θ_{act} is the actual throughput of the assembly system, computed by considering the assembly system to be a Markov process and calculating its steady state probability distribution. θ_{Lip} stands for the approximate value for the throughput as computed by Lipper and Sengupta [14], and θ_{apr} stands for the approximate value we are suggesting, i.e. $\theta_{\text{apr}} = (\theta_{\text{ub}} + \theta_{\text{hlb}})/2$.

From table 1 it is apparent that our approximation does better than the approximation given in Lipper and Sengupta [14] in some instances. Our approach also yields bounds, since θ_{ub} is a guaranteed upper bound and θ_{hlb} seems to consistently underestimate the throughput (at least in all the computations we have done). In addition, our approximation is computationally very simple, which allows it to be used in routines to optimize the system performance with respect to variable system parameters. The closed form expressions given here are potentially useful as the basis for determining optimal buffer sizes.

However, as pointed out earlier, our method works only for two inputs, whereas the approach of Lipper and Sengupta can handle more than two inputs to the assembly machine. Further work is needed to develop simple closed-form approximations for the case with more than two inputs and to refine the bounds and heuristics for average inventories.

Appendix A

COMPUTING θ_{lb2}

To compute θ_{lb2}, we need to compute $E[\max\{\text{Erlang}(\lambda_1, n), \text{Erlang}(\lambda_2, n), \text{Erlang}(\mu, n)\}]$, where the three random variables are mutually independent.

Define X_m, $m = 1, 2, 3\ldots$ to be independent and exponentially distributed random variables, each with parameter λ_1. Similarly define Y_m to be independent and exponentially distributed random variables with parameter λ_2, and Z_m to be independent and exponentially distributed random variables with parameter μ. Let

$$X^{(i)} = \sum_{m=1}^{i} X_m,\ Y^{(j)} = \sum_{m=1}^{j} Y_m,\ Z^{(k)} = \sum_{m=1}^{k} Z_m.$$

Also define $T(i, j, k) = E[\max\{X^{(i)}, Y^{(j)}, Z^{(k)}\}]$. Using this definition, our aim is to compute $T(n, n, n)$.

If $i > 0$, $j > 0$ and $k > 0$, by conditioning on $\min\{X_1, Y_1, Z_1\}$, we can write

$$T(i, j, k) = \frac{1}{\lambda_1 + \lambda_2 + \mu} + \frac{\lambda_1}{\lambda_1 + \lambda_2 + \mu} T(i-1, j, k) + \frac{\lambda_2}{\lambda_1 + \lambda_2 + \mu} T(i, j-1, k) + \frac{\mu}{\lambda_1 + \lambda_2 + \mu} T(i, j, k-1).$$

Extending this to the general case, we can write

$$\begin{aligned} T(i, j, k) = {} & I_{\{i>0, j>0, k>0\}} \left\{ \frac{1}{\lambda_1 + \lambda_2 + \mu} + \frac{\lambda_1}{\lambda_1 + \lambda_2 + \mu} T(i-1, j, k) \right. \\ & \left. + \frac{\lambda_2}{\lambda_1 + \lambda_2 + \mu} T(i, j-1, k) + \frac{\mu}{\lambda_1 + \lambda_2 + \mu} T(i, j, k-1) \right\} \\ & + I_{\{i>0, j>0, k=0\}} \left\{ \frac{1}{\lambda_1 + \lambda_2} + \frac{\lambda_1}{\lambda_1 + \lambda_2} T(i-1, j, 0) \right. \\ & \left. + \frac{\lambda_2}{\lambda_1 + \lambda_2} T(i, j-1, 0) \right\} \\ & + I_{\{i>0, j=0, k>0\}} \left\{ \frac{1}{\lambda_1 + \mu} + \frac{\lambda_1}{\lambda_1 + \mu} T(i-1, 0, k) \right. \\ & \left. + \frac{\mu}{\lambda_1 + \mu} T(i, 0, k-1) \right\} \\ & + I_{\{i=0, j>0, k>0\}} \left\{ \frac{1}{\lambda_2 + \mu} + \frac{\lambda_2}{\lambda_2 + \mu} T(0, j-1, k) \right. \\ & \left. + \frac{\mu}{\lambda_2 + \mu} T(0, j, k-1) \right\} \\ & + I_{\{i>0, j=0, k=0\}} \left\{ \frac{1}{\lambda_1} + T(i-1, 0, 0) \right\} \\ & + I_{\{i=0, j>0, k=0\}} \left\{ \frac{1}{\lambda_2} + T(0, j-1, 0) \right\} \\ & + I_{\{i=0, j=0, k>0\}} \left\{ \frac{1}{\mu} + T(0, 0, k-1) \right\}. \end{aligned}$$

Given that $T(0, 0, 0) = 0$, $T(1, 0, 0) = 1/\lambda_1$, $T(0, 1, 0) = 1/\lambda_2$ and $T(0, 0, 1) = 1/\mu$, we can compute any $T(i, j, k)$ iteratively. Thus we can compute $T(n, n, n)$.

Appendix B

COMPUTING $\mathscr{I}_{\text{lb1}c}$ AND $\mathscr{I}_{\text{lb2}c}$

We are given an assembly system $\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ with $\mu = \infty$. Clearly $N_1(t)$ and $N_2(t)$ cannot both be non-zero at the same time. Hence we can denote the state $(N_1(t), N_2(t))$ using a single variable $N(t)$, where

$$N(t) > 0 \Leftrightarrow N_1(t) = N(t),\ N_2(t) = 0$$

$$N(t) < 0 \Leftrightarrow N_1(t) = 0,\ N_2(t) = -N(t) \text{ and}$$

$$N(t) = 0 \Leftrightarrow N_1(t) = N_2(t) = 0.$$

Thus we have a Markov process $\{N(t);\ t \geqslant 0\}$ on the state space $\{-K_2,\ K_2 + 1, \ldots, -1,\ 0,\ 1, \ldots, K_1 - 1,\ K_1\}$. Its steady state probabilities $\{\pi(k);\ k = -K_2, \ldots, K_1\}$ are given by

$$\pi(k) = \pi(0)\rho^k$$

where

$$\rho = \lambda_1/\lambda_2, \text{ and } \pi(0) = \left[\sum_{k=-K_2}^{K_1} \rho^k\right]^{-1}.$$

Now

$$\mathscr{I}_{\text{lb1}c} = \sum_{k=1}^{K_1} k\pi(k) = \begin{cases} \dfrac{\rho(1-\rho^{K_1}) - K_1\rho^{K_1+1}(1-\rho)}{(1-\rho)^2} & \text{if } \rho \neq 1 \\ K_1(K_1+1)/2 & \text{if } \rho = 1. \end{cases}$$

$\mathscr{I}_{\text{lb2}c}$ is obtained from the above expression by replacing K_1 by K_2 and ρ by $\sigma = 1/\rho$.

Acknowledgement

The authors would like to express their gratitude to an annonymous referee for several useful suggestions for improving the content and presentation of this paper.

References

[1] T. Altiok, Approximate analysis of exponential tandem queues with blocking, European Journal of Operations Research 11 (1982) 390.

[2] M.H. Ammar, Modelling and analysis of unreliable manufacturing assembly networks with finite storage, MIT Laboratory for Information and Decision Sciences, Report LIDS-TH-1004, June 1980.
[3] U.N. Bhat, Finite capacity assembly-like queues, Queueing Systems: Theory and Applications 1 (1986) 85.
[4] F. Bonomi, An approximate analysis for a class of assembly-like queues, Queueing Systems: Theory and Applications 1 (1987) 289.
[5] J.A. Buzacott, Automatic transfer lines with buffer stocks, International Journal of Production Research 5 (1967) 183.
[6] J.B. Dennis, Data flow super computers, IEEE Computer 13, 11 (1980) 48.
[7] S.B. Gershwin and I.C. Schick, Modelling and analysis of three-stage transfer lines with unreliable machines and finite buffers, Operations Research 31 (1983) 354.
[8] D. Gross and C.M. Harris, *Fundamentals of Queueing Theory* (John Wiley & Sons, New York, 1985).
[9] J.M. Harrison, Assembly-like Queues, Journal of Applied Probability 10 (1973) 354.
[10] J.M. Hatcher, The effect of internal storage on the production rate of a series having exponential service times, AIIE Transactions 1 (1969) 150.
[11] F.S. Hillier and R.N. Boling, Finite queues in series with exponential or Erlang service times – A numerical approach, Operations Research 15 (1967) 286.
[12] G.C. Hunt, Sequential arrays of waiting lines, Operations Research 4 (1956) 674.
[13] G. Latouche, Queues with paired customers, Journal of Applied Probability 18 (1981) 684.
[14] E.H. Lipper and B. Sengupta, Assembly-like queues with finite capacity: bounds, asymptotics and approximations, Queueing Systems: Theory and Applications 1 (1986) 67.
[15] Y. Monden, Adaptable Kanban system helps Toyota maintain just-in-time production, Industrial Engineering 13 (1981) 29.
[16] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models* (The Johns Hopkins University Press, Baltimore, 1981).
[17] S.M. Ross, *Stochastic Processes* (John Wiley & Sons, New York, 1983).