Queueing Systems 17 (1994) 471–490

Kitting process in a stochastic assembly system

Pradip Som, W.E. Wilhelm and R.L. Disney

Department of Industrial Engineering, Texas A&M University, College Station, TX 77843-3131, USA

Received 14 March 1994

In small-lot, multi-product, multi-level assembly systems, kitting (or accumulating) components required for assembly plays a crucial role in determining system performance, especially when the system operates in a stochastic environment. This paper analyzes the kitting process of a stochastic assembly system, treating it as an assembly-like queue. If components arrive according to Poisson processes, we show that the output stream departing the kitting operation is a Markov renewal process. The distribution of time between kit completions is also derived. Under the special condition of identical component arrival streams having the same Poisson parameter, we show that the output stream of kits approximates a Poisson process with parameter equal to that of the input stream. This approximately decouples assembly from kitting, allowing the assembly operation to be analyzed separately.

Keywords: Kitting, assembly, Markov renewal process, double-ended queue.

1. Introduction

Traditionally, material flow analysis in assembly systems has been based on the assumption that the system operates deterministically. In recent years, attention has been directed to a more realistic analysis of assembly systems, explicitly treating the stochastic events that influence operations. An important aspect of assembly operations is kitting (or accumulating) required components and releasing the kit to initiate assembly. Due to the stochastic nature of component availability, stock-outs often occur in component inventories, thereby disrupting kitting and, consequently, assembly schedules. The goal of this paper is to better understand the kitting process in a stochastic assembly system, which we treat as an assembly-like queue.

This paper models the kitting process of an assembly system as a Markov renewal process, assuming that component arrival streams follow independent Poisson distributions. The assembly system is assumed to have a structure similar to that described in Hopp and Simon [7] and is shown in figure 1.

P_1 and P_2 are machines that process components (to prepare them for assembly) and P_3 is the assembly machine. I_1 and I_2 are the buffers for components, I_0 is the buffer for kits, and I_3 is the buffer for the end-product. P_1 and

© J.C. Baltzer AG, Science Publishers

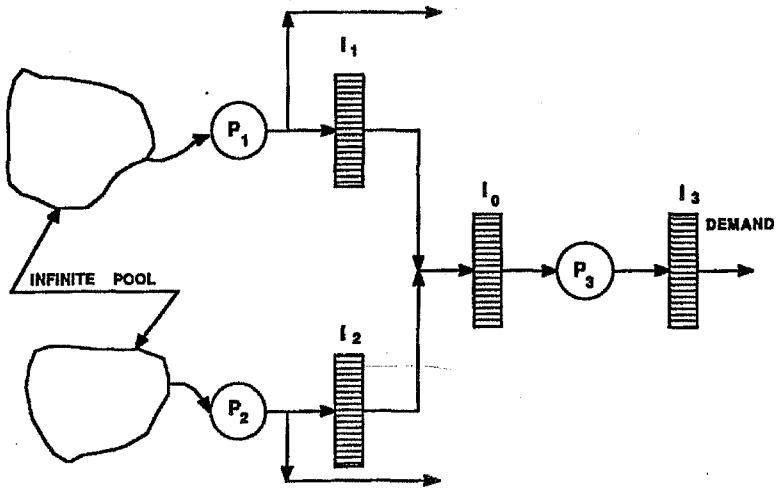

Fig. 1. Stochastic assembly system.

P_2 work independently, withdraw raw materials from their respective pools of unlimited supply, and deliver processed components to buffers I_1 and I_2, respectively. A component arriving at buffer I_1 (I_2) is immediately kitted with a part from buffer I_2 (I_1) if one is available, and a "kit" is said to be composed. If a kit cannot be composed, the processed part is held in buffer I_1 (I_2) to await the arrival of a "matching" part at buffer I_2 (I_1). Once composed, a kit matching components from I_1 and I_2 is sent immediately to I_0 and the kit is considered to be one arrival at I_0. If the arriving kit finds I_0 empty and P_3 idle, it is immediately placed in the assembly machine P_3. Otherwise, the kit is held in buffer I_0.

We assume that buffers of components have limited capacity and that each component is processed according to an exponential distribution (before kitting) to prepare it for assembly. When P_3 completes an assembly, it withdraws a kit (i.e. two matched components) from I_0, whenever available, then assembles another end product and delivers it to buffer I_3. If a kit is not available in I_0 when P_3 completes an assembly, it remains idle until a completed kit arrives. Demands for end products arrive at I_3; each demand is assumed to be for a lot of unit size and is satisfied immediately if stock is available. Unsatisfied demands are backordered, causing the inventory position at I_3 to take on negative values.

Our primary result is to show that the output stream departing the kitting operation is a Markov renewal process. In the special case in which component arrival streams have the same Poisson parameter, we are able to show that the output stream approximates a Poisson process with parameter the same as that of the arrival streams.

Regarding the modus operandi of the assembly system, Harrison [6] showed that a sufficient condition for stability of operations of such systems is that component buffer sizes be finite. For a system with finite buffers, we show that, in the long run, the probability of observing inventory position j at I_1 (I_2) depends on

the inventory position j. Also, considering the special case of component arrival streams with the same Poisson parameter, we show that the kit completion process well approximates a Poisson process when the component buffers are large enough, permitting the kitting and assembly operation to be decoupled so that downstream operations can be analyzed separately.

Stochastic assembly systems are often studied as assembly-like queues. Harrison [6] showed that an assembly system with input streams that are independent renewal processes and with no inventory capacity limitations for any stream are unstable. He also showed that, under these conditions, the limiting distribution of the time that parts wait for assembly converges to a defective distribution.

Since we assume that two components are required to compose a kit, the queues of components form a double-ended queue [5, 8]. A double-ended queue can be best described by the well known taxi-cab problem where taxis and passengers form two different queues. A customer waits in its queue and leaves it as soon as a taxi is available; taxis wait in queue for customers and leave when a customer is available. The two queues are interdependent and their combination is known as a double-ended queue where it is known that the related queueing process is a random walk on $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$ and is transient or null unless the queues are bounded. The kitting process under study can be considered as a double-ended queue of the type examined by Kashyap and Chaudhury [9]. They showed that each queue length distribution is independent of occupancy when arrival rates to the double-ended queue are equal. They also derived the distribution of waiting times in double-ended queues but made no attempt to analyze its output process.

Bhat [2] incorporated limited buffer capacities in assembly like queues and derived expressions for the stationary probability vector of the queue length. Latouche [10] considered assembly systems with Poisson procurement processes and exponential processing times and derived conditions required for stability. Assembly networks that represent one-time production (for example, space-shuttle, aircraft prototype, etc.) are analyzed by Saboo and Wilhelm [11] and Wilhelm et al. [13].

The output processes from queues operating according to various disciplines are reviewed by Disney and Konig [4] in detail. They describe the characteristics of the output processes resulting from $GI/D/s$, $M/M/s$, $M/GI/1/L$, $M/E_k/1/L$, $M/GI/\infty$, $GI/GI/1/L$ and $GI/M/1/L$ systems. Apparently, the output process of a double-ended queue has not been studied previously. In this paper we analyze such a process as a part of our study of the kitting process.

We have organized this paper in five sections. The fundamentals and pertinent assumptions are presented in section 2. Section 3 relates the formulation of a Markov renewal process which describes the kitting operation. The model is evaluated in section 4 by determining the state transition matrix P, the time-stationary probability vector Π, and the distribution of time between kit

completions, which is shown to be approximately Poisson under certain conditions. Practical implications of analytical results are described and conclusions are presented in section 5.

2. Fundamentals

The structure of the assembly system under analysis is presented in figure 1. A little thought indicates that it is not possible for both buffers I_1 and I_2 to have positive stock levels at the same time. An arrival which increases the stock level of one of the buffers to a positive value creates a "virtual backorder" at the other buffer. At any time t ($t > 0$), the inventory position "M" (defined as the number of parts on hand plus on order minus the number on back order) on one buffer is associated with inventory position "$-M$" in the other, and equality holds only when the inventory position is zero (0) for both buffers I_1 and I_2. The inventory positions at I_1 and I_2 may thus be viewed as "mirror images" of one another, a special structure which we exploit to analyze the kitting process.

Since the purpose of this paper is to characterize the kitting process, we study the stream of arrivals to I_0 (i.e., the output of the kitting process) in the following sections and ignore the process downstream of I_0. We present a thorough analysis of the downstream assembly system in a companion paper (Som and Wilhelm [12]).

Our model, which is based on the structure described in this section, relies upon three fundamental assumptions:

(i) Processing times at the part processing machines, P_1 and P_2, are independent, identically distributed, non-negative exponential random variables with rates μ_1 and μ_2, respectively.

(ii) The capacities of buffers I_1 and I_2 are bounded from above by K_1 and K_2, respectively, representing practical limitations on buffer space, and, according to Harrison [6], allowing the system to reach a steady state. No capacity restriction is imposed on I_0.

(iii) P_1 (P_2) prepares parts exclusively for I_1 (I_2). However, when I_1 (I_2) is fitted to capacity K_1 (K_2), additional arrivals are not processed in the system under analysis (e.g., they may be processed and assembled by a subcontractor).

In the following sections we formulate the model and analyze it as a Markov renewal process.

3. Formulation of a Markov renewal process

The inventory positions at I_1 and I_2 change with the arrival and departure of components to and from the respective buffers. We define the *mirror image process*

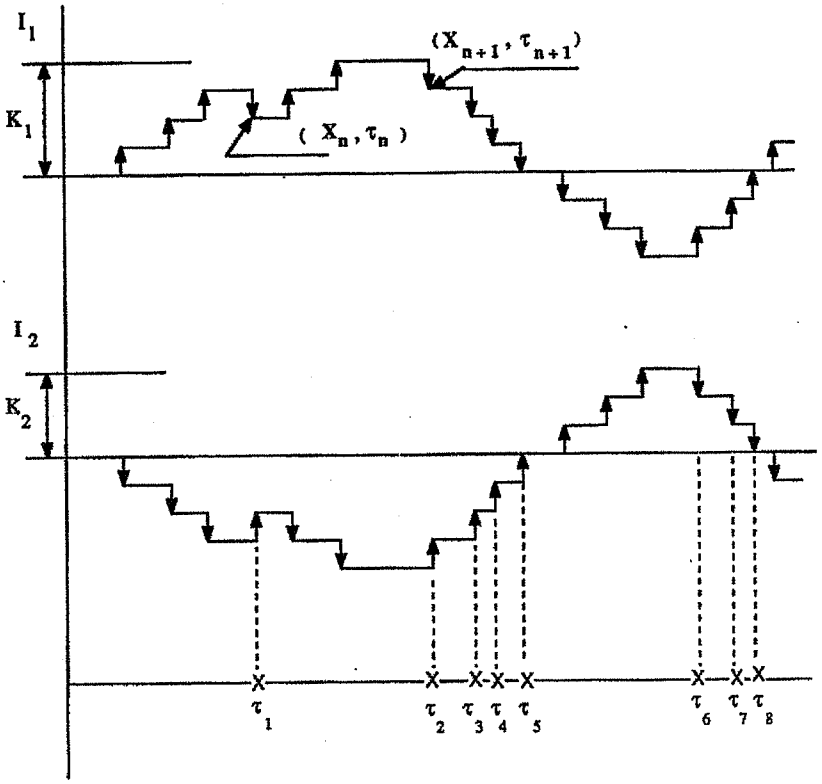

Fig. 2. Mirror image process and output process.

(X, T) as a *marked point process*, which characterizes the inventory positions or states at the arrival and departure epochs. The sample path diagram of the *mirror image process* is presented in figure 2.

Thus,

$$(X, T) = \{X_m, T_m : m \in \mathbb{N}\},$$

in which,

$$X_m = \{{}^1X_m, {}^2X_m\},$$

T_m = time of mth state change epoch,

${}^1X_m$ = inventory position at buffer I_1 at time T_m,

${}^2X_m$ = inventory position at buffer I_2 at time T_m.

Due to the mirror image property of the inventory positions at I_1 and I_2, at any random time T_m, ${}^1X_m = {}^1x_m$ implies ${}^2X_m = -{}^1x_m$; or, equivalently, ${}^2X_m = {}^2x_m$ implies ${}^1X_m = -{}^2x_m$. Hence, it is obvious that the Mirror Image Process may be analyzed by viewing the inventory position just at I_1 (or, equivalently, just at I_2).

Whenever matching components are available at buffers I_1 and I_2, a kit is composed (instantaneously) and sent to I_0. These departure epochs (occurring simultaneously from both I_1 and I_2) and the corresponding inventory position at I_1 describe another *marked point process* which we define as the *output process.* By observing the inventory at I_1, it is apparent that a particular subset of the epochs $\{T_m : m \in \mathbb{N}\}$, marked by a decrease in the positive inventory position or an increase in the negative inventory position, constitutes kit completion as well as state change epochs in the *output process.*

These output epochs are a sub-sequence of the sequence $\{T_m : m \in \mathbb{N}\}$, defined as $\tau = \{\tau_n : n \in \mathbb{N}\}$ with $0 = \tau_0 \leq \tau_1 \leq \tau_2 \leq \tau_3 \leq \ldots$ such that for $\omega \in \Omega$, $\tau_0(\omega) = T_0(\omega) = 1$, $\tau_n(\omega) = T_k(\omega)$, $n \geq 1$, in which $k = \min\{m \in \mathbb{N} : n \leq \sum_{j=1}^{m} \mathbf{1}_{\{|{}^1X_{j-1}| > |{}^1X_j|\}}(\omega)\}$ and $\mathbf{1}_{\{\mathbf{X}\}}(\cdot)$ is an indicator function. Define $D_{n+1} = \tau_{n+1} - \tau_n$ as the time between successive departures, n and $n+1$. For $n \in \mathbb{N}$, the random variable $D_n : \Omega \to \mathbb{R}^+$ represents the length of the nth inter-departure interval. Then $\tau_{n+1} = \tau_n + D_{n+1}$, $n \in \mathbb{N}$, defines the time of the $(n+1)$th departure. The set $\tau = \{\tau_n : n \in \mathbb{N}\}$ defines the *output time process.*

For each $n \in \mathbb{N}$, define the random variable $Z_n : \Omega \to E$ as the inventory position at the buffer I_1 or the system state of the *output process* immediately after the nth departure epoch τ_n. The set $Z = \{Z_n : n \in \mathbb{N}\}$ defines the *output state process,* and the joint random variables $\{Z, \tau\} = \{Z_n, \tau_n : n \in \mathbb{N}\}$ define the *output process.* Here, D_n depends on the present state Z_n and the next state Z_{n+1}. However, given these states, D_n is independent of previous D_k and Z_k for $k = 1, \ldots, n-1$, indicating that the *output process* $\{Z, \tau\}$ is a Markov renewal process on the state space E. Since a Markov renewal process is completely characterized by its semi-Markov kernel $Q(i,j,t)$, we study this kernel in the following subsection.

DETERMINATION OF THE SEMI-MARKOV KERNEL $Q(i,j,t)$

The semi-Markov kernel of the *output process* $\{Z, \tau\}$ may be expressed as

$$Q(i,j,t) = Pr\{Z_{n+1} = j, \tau_{n+1} - \tau_n \leq t | Z_n = i\}.$$

For convenience, the semi-Markov kernel is expressed in the Laplace transform domain as $L\{Q(i,j,\mathrm{d}t)\} = Q\{i,j,\mathrm{d}s\}$.

The Laplace transform of $(\mathrm{d}/\mathrm{d}t)Pr\{Z_{n+1} = j, \tau_{n+1} - \tau_n \leq t | Z_n = i\}$, expressed as $L[\mathrm{d}P\{Z_{n+1} = j, \tau_{n+1} - \tau_n \leq t | Z_n = i\}]$, can be shown to have five different forms, depending upon inventory positions at epochs τ_n and τ_{n+1}. We describe the five cases below.

Case I. The starting (i.e., at τ_n) inventory position is non-negative and it *does not* reach the positive boundary K_1 before the time of the next departure (i.e., at τ_{n+1}).

Certain combinations of i and j define case I:

(i) $0 < i \leq K_1 - 1, \quad i - 1 \leq j \leq K_1 - 2$, and

(ii) $i = 0, \quad 0 < j \leq K_1 - 2.$

Then,

$$\mathrm{d}P\{Z_{n+1} = j, \tau_{n+1} - \tau_n \leq t | Z_n = i\} = \frac{\mathrm{e}^{-\mu_1 t}(\mu_1 t)^{j-i+1}}{(j-i+1)!}\mu_2 \mathrm{e}^{-\mu_2 t}\,\mathrm{d}t. \tag{1}$$

Since we are looking at two consecutive kit completion epochs, τ_n and τ_{n+1}, at which inventory positions at I_1 are i and j respectively, $j - i + 1$ components must have arrived at I_1 before any arrival at I_2.

In Laplace transform form,

$$L[\mathrm{d}P\{Z_{n+1} = j, \tau_{n+1} - \tau_n \leq t | Z_n = i\}] = \left(\frac{\mu_2}{\mu_1}\right)\left(\frac{\mu_1}{\mu_1 + \mu_2}\right)^{j-i+2}\left(\frac{\mu_1 + \mu_2}{\mu_1 + \mu_2 + s}\right)^{j-i+2}. \tag{2}$$

The other four cases follow similarly.

Case II.

(i) $-K_2 + 1 \leq i < 0, \quad -K_2 + 2 \leq j \leq i + 1$, and

(ii) $i = 0, \quad -K_2 + 2 \leq j < 0.$

$$L[\mathrm{d}P\{Z_{n+1} = j, \tau_{n+1} - \tau_n \leq t | Z_n = i\}] = \left(\frac{\mu_1}{\mu_2}\right)\left(\frac{\mu_2}{\mu_1 + \mu_2}\right)^{|j|-|i|+2}\left[\frac{\mu_1 + \mu_2}{\mu_1 + \mu_2 + s}\right]^{|j|-|i|+2}. \tag{3}$$

Case III.

$0 \leq i \leq K_1 - 1, \quad j = K_1 - 1.$

$$L[\mathrm{d}P\{Z_{n+1} = K_1 - 1, \tau_{n+1} - \tau_n \leq t | Z_n = i\}] = \left(\frac{\mu_1}{\mu_1 + \mu_2}\right)^{K_1 - i}\left[\frac{\mu_2}{\mu_2 + s}\right]\left[\frac{\mu_1 + \mu_2}{\mu_1 + \mu_2 + s}\right]^{K_1 - i}. \tag{4}$$

Case IV.

$-K_2 + 1 \leq i \leq 0, \quad j = -K_2 + 1.$

$$L[\mathrm{d}P\{Z_{n+1} = -K_2 + 1, \tau_{n+1} - \tau_n \leq t | Z_n = i\}] = \left(\frac{\mu_2}{\mu_1 + \mu_2}\right)^{K_2 + i}\left[\frac{\mu_1}{\mu_1 + s}\right]\left[\frac{\mu_1 + \mu_2}{\mu_1 + \mu_2 + s}\right]^{K_2 + i}. \tag{5}$$

Table 1
Semi-Markov kernel $L\{Q(i,j,\mathrm{d}s)\}$ – unequal arrival rates.

$L\{Q(i,j,\mathrm{d}s)\}$

$$
= \begin{array}{c|cccccccccccc}
 & -K_2+1 & -K_2+2 & \cdots & -2 & -1 & 0 & 1 & 2 & \cdots & K_1-2 & K_1-1 \\
\hline
-K_2+1 & \vdots & \vdots & \vdots & & & & & & & & \\
-K_2+2 & & & & & & & & & & & \\
\vdots & & & & & & & 0 & & & & \\
 & b_1[(1-\nu)a]^{K_1+i} & & & & & & & & & & \\
-2 & \vdots & \vdots & \rho[(1-\nu)a]^{|j|-|i|+2} & \cdots & & & & & & & \\
-1 & & & \vdots & \cdots & \rho(1-\nu)a & & & & & & \\
0 & & & & & 2\nu(1-\nu)a^2 & \cdots & & & & & \\
1 & & & & & \frac{1}{\rho}\nu a & \cdots & & & & & \\
2 & & & & & & \cdots & & & & & \\
\vdots & & 0 & & & & \cdots & \frac{1}{\rho}[a\nu]^{|j|-|i|+2} & \vdots & & & \\
K_1-2 & & & & & & & \vdots & & & b_2[\nu a]^{K_1-i} & \\
K_1-1 & & & & & & & & & & \vdots &
\end{array}, \quad (7)
$$

$$
\rho=\frac{\mu_1}{\mu_2}, \quad \nu=\frac{\mu_1}{\mu_1+\mu_2}, \quad a=\frac{\mu_1+\mu_2}{\mu_1+\mu_2+s}, \quad b_1=\frac{\mu_1}{\mu_1+s}, \quad b_2=\frac{\mu_2}{\mu_2+s}.
$$

Table 2
Transition probability matrix $P(i,j)$ – unequal arrival rates.

$$P(i,j) = \begin{array}{c} \\ -K_2+1 \\ -K_2+2 \\ \vdots \\ \\ -2 \\ -1 \\ 0 \\ 1 \\ 2 \\ \vdots \\ \\ K_1-2 \\ K_1-1 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} -K_2+1 & -K_2+2 & \cdots & -2 & -1 & 0 & 1 & 2 & \cdots & K_1-2 & K_1-1 \end{array} \\ \begin{bmatrix} \vdots & \vdots & \vdots & & & & & & & & \\ & & & & & & 0 & & & & \\ (1-\nu)^{K_2+i} & & & & & & & & & & \\ \vdots & \vdots & \rho(1-\nu)^{|j|-|i|+2} & \cdots & & & & & & & \\ & & \vdots & \cdots & \rho(1-\nu) & & & & & & \\ & & & & 2\nu(1-\nu) & \cdots & & & & & \\ & & & & \frac{1}{\rho}\nu & \cdots & & & & & \\ & & & & & \cdots & & & & & \\ & 0 & & & & \cdots & \frac{1}{\rho}\nu^{|j|-|i|+2} & \vdots & & & \\ & & & & & & \vdots & & & \nu^{K_1-i} & \\ & & & & & & & & & \vdots & \end{bmatrix} \end{array}, \quad (8)$$

$$\rho = \frac{\mu_1}{\mu_2}, \quad \nu = \frac{\mu_1}{\mu_1+\mu_2}.$$

Case V.

$i = 0, \quad j = 0.$

$$L[\mathrm{d}P\{Z_{n+1} = 0, \tau_{n+1} - \tau_n \leq t | Z_n = 0\}]$$

$$= \frac{2\mu_1\mu_2}{(\mu_1+\mu_2)^2}\left[\frac{\mu_1+\mu_2}{\mu_1+\mu_2+s}\right]^2. \tag{6}$$

The interval $\tau_{n+1} - \tau_n$ includes an initial period which has an exponential distribution with rate $\mu_1 + \mu_2$. Using Bernoulli probabilities $\mu_1/(\mu_1+\mu_2)$ and $\mu_2/(\mu_1+\mu_2)$ and convolving with the distribution of the remainder of the interval, we get the above result.

Combining equations (2) through (6), we obtain the semi-Markov kernel $Q(i, j, t)$, which is expressed in Laplace transform form and is presented as equation (7) in table 1. The state transition matrix P of the underlying Markov chain Z embedded at time τ_n is obtained by setting $s = 0$ in equation (7) and is presented as equation (8) in table 2. An analysis of the output process $\{Z, \tau\}$ is presented in the following section.

4. Analysis

In this section, we analyze the *output process* $\{Z, \tau\}$ deriving the following:

(i) the stationary probability vector Π of the underlying Markov chain Z, and

(ii) the distribution of time between kit completions.

The vector Π indicates the time-stationary probably distribution of the inventory position at I_1, observed at a randomly selected kit completion epoch.

DETERMINATION OF STATIONARY PROBABILITY VECTOR Π

Clearly, the output process $\{Z, \tau\} = \{Z_n, \tau_n : n \in \mathbb{N}\}$ is an irreducible, nonnull, recurrent, and persistent Markov renewal process for $K_1, K_2 < \infty$; under these conditions, it possesses a stationary distribution defined as Π [3]. Note that the process $\{Z, \tau\}$ will be recurrent null, if K_1 and K_2 are infinite. The stationary probability vector Π of the underlying Markov chain is obtained from the set of equations expressed in the matrix form

$$\Pi = \Pi P.$$

Using equation (14) for P, the balance equations can be expressed for specific states $-K_2+1 \le j \le K_1-1$ as

$$\Pi(0) = \Pi(-1)\rho(1-\nu) + \Pi(0)2\nu(1-\nu) + \Pi(1)(1/\rho)\nu, \tag{9}$$

$$\Pi(K_1-1) = \rho\Pi(K_1-2), \tag{10}$$

$$\Pi(-K_2+1) = (1/\rho)\Pi(-K_2+2), \tag{11}$$

$$\Pi(j) = \Pi(0)(1/\rho)\nu^{(j+2)} + \Pi(1)(1/\rho)\nu^{(j+1)} + \Pi(2)(1/\rho)\nu^{j} + \ldots + (1/\rho)\Pi(j+1), \quad j = 1,2,\ldots,K_2-2, \tag{12}$$

$$\Pi(j) = \Pi(0)\rho(1-\nu)^{(-j+2)} + \Pi(1)\rho(1-\nu)^{(-j+1)} + \Pi(2)\rho(1-\nu)^{(-j)} + \ldots + \rho(1-\nu)\Pi(-j+1), \quad j = -1,-2,\ldots,-K_2+2, \tag{13}$$

in which

$$\rho = \mu_1/\mu_2, \quad \nu = \mu_1/(\mu_1+\mu_2).$$

In addition, we have the normalizing expression

$$\sum_j \Pi(j) = 1. \tag{14}$$

The solution to equations (9)–(14) can be expressed as

$$\Pi(0) = \frac{(\rho-1)(\mu_1+\mu_2)}{\mu_2(\rho^{K_1}-\rho^{-K_2})}, \tag{15}$$

$$\Pi(j) = \nu\rho^j\Pi(0), \qquad j = 1,2,\ldots,K_1-1, \tag{16}$$

$$\Pi(j) = (1-\nu)\rho^j\Pi(0), \quad j = -1,-2,\ldots,-K_2+1. \tag{17}$$

It may be observed that $\Pi(j)$, the stationary probability of positive (negative) stock in buffer I_1 observed at a kit completion time, depends on the stock position, j.

DISTRIBUTION OF TIME BETWEEN KIT COMPLETIONS, D_n

To determine the distribution of time between kit completions, we concentrate on analyzing the *output time process* $\tau = \{\tau_n : n \in \mathbb{N}\}$, which specifies the arrival stream (of kits) to buffer I_0.

Considering the stationary distribution Π of the underlying Markov chain Z and for $t \in \mathbb{R}^+$, the distribution of time between two consecutive kit completions is

given by

$$P\{\tau_{n+1} - \tau_n \leq t\} = \Pi Q(i, j, t)U, \tag{18}$$

in which U is a column vector with all elements equal to 1.

Expressing equation (18) in Laplace transform form we obtain:

$$L[\mathrm{d}P(\tau_{n+1} - \tau_n \leq t)] = \Pi Q(i, j, \mathrm{d}s)U. \tag{19}$$

Substituting the values of Π and $Q(i, j, \mathrm{d}s)$ from equations (15) to (17) and (7) into equation (19),

$$\begin{aligned} L[\mathrm{d}P(\tau_{n+1} - \tau_n \leq t)] = {} & \left(\frac{\mu_1}{\mu_1 + s}\right)\left[\Pi(0) + \sum_{j=-1}^{-K_2+1}(1-\nu)\rho^j\Pi(0)\right] \\ & + \left(\frac{\mu_2}{\mu_2 + s}\right)\left[\Pi(0) + \sum_{j=1}^{K_1-2}\nu\rho^j\Pi(0)\right] \\ & - \Pi(0)\left(\frac{\mu_1 + \mu_2}{\mu_1 + \mu_2 + s}\right). \end{aligned} \tag{20}$$

It is apparent that if equation (20) is inverted (i.e., to be the time domain), the distribution of the time between kit completions, D_{n+1}, would be the weighted sum of three exponential distributions with rates μ_1, μ_2 and $\mu_1 + \mu_2$.

A SPECIAL CASE WITH $\mu_1 = \mu_2 = \mu$

This section specializes the case in which component processing times at machines P_1 and P_2 are independent exponential random variables with the same rates (i.e., $\mu_1 = \mu_2 = \mu$). In practice, this situation may occur when components are obtained from independent suppliers with identical (and independent) lead time distributions. Also, the same situation may occur during "in-house" production where the machines employed, P_1 and P_2, are identical (and independent). In the following sub-sections we show that the distribution of time between kit completions, D_n, can be approximated by independent and identically distributed exponential random variables.

APPROXIMATION OF D_n BY THE EXPONENTIAL DISTRIBUTION

Making appropriate changes in equations (7) and (8) to accommodate the special case, the semi-Markov kernel, $Q(i, j, t)$, Laplace transform form and the

Table 3
Semi-Markov kernel $L\{Q(i,j,\mathrm{d}s)\}$ – equal arrival rates.

$$L\{Q(i,j,\mathrm{d}s)\} = \begin{array}{c} \\ -K_2+1 \\ -K_2+2 \\ \vdots \\ \\ -2 \\ -1 \\ 0 \\ 1 \\ 2 \\ \vdots \\ K_1-2 \\ K_1-1 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} -K_2+1 & -K_2+2 & \cdots & -2 & -1 & 0 & 1 & 2 & \cdots & K_1-2 & K_1-1 \end{array} \\ \begin{bmatrix} \vdots & \vdots & \vdots & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & 0 & & & & \\ b_1(\frac{1}{2}a)^{K_1+i} & & & & & & & & & & \\ \vdots & \vdots & (\frac{1}{2}a)^{|j|-|i|+2} & \cdots & & & & & & & \\ & & \vdots & \cdots & \frac{1}{2}a & & & & & & \\ & & & & (\frac{1}{2}a)^2 & \cdots & & & & & \\ & & & & \frac{1}{2}a & \cdots & & & & & \\ & & & & & \cdots & & & & & \\ & 0 & & & & \cdots & (\frac{1}{2}a)^{|j|-|i|+2} & \vdots & & & \\ & & & & & & \vdots & & & b_2(\frac{1}{2}a)^{K_1-i} & \\ & & & & & & & & & \vdots & \end{bmatrix} \end{array}, \quad (21)$$

$$a = \frac{\mu_1+\mu_2}{\mu_1+\mu_2+s}, \quad b_1 = \frac{\mu_1}{\mu_1+s}, \quad b_2 = \frac{\mu_2}{\mu_2+s}.$$

Table 4
Transition probability matrix $P(i, j)$ – equal arrival rates.

$$P(i,j) = \begin{array}{c} \\ -K_2+1 \\ -K_2+2 \\ \vdots \\ \\ -2 \\ -1 \\ 0 \\ 1 \\ 2 \\ \vdots \\ K_1-2 \\ K_1-1 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} -K_2+1 & -K_2+2 & \cdots & -2 & -1 & 0 & 1 & 2 & \cdots & K_1-2 & K_1-1 \end{array} \\ \left[\begin{array}{cccccccccccc} \vdots & \vdots & \vdots & & & & & & & & \\ & & & & & & 0 & & & & \\ \left(\frac{1}{2}\right)^{K_1+i} & & & & & & & & & & \\ \vdots & \vdots & \left(\frac{1}{2}\right)^{|j|-|i|+2} & \cdots & & & & & & & \\ & & \vdots & \cdots & \frac{1}{2} & & & & & & \\ & & & & \frac{1}{2} & \cdots & & & & & \\ & & & & \frac{1}{2} & \cdots & & & & & \\ & & & & & \cdots & & & & & \\ & 0 & & & & \cdots & \left(\frac{1}{2}\right)^{|j|-|i|+2} & \vdots & & & \\ & & & & & & \vdots & & \left(\frac{1}{2}\right)^{K_1-i} & & \\ & & & & & & & & \vdots & & \end{array}\right] \end{array}, \quad (22)$$

transition probability matrix P of the underlying Markov chain Z may be expressed by equations (21) and (22) which are presented in tables 3 and 4, respectively.

The stationary probabilities of this Markov chain are given by

$$\Pi(0) = \frac{2}{K_1 + K_2}, \tag{23}$$

$$\Pi(j) = \frac{1}{K_1 + K_2}, \quad \forall j \neq 0. \tag{24}$$

These results have striking similarities – but at the same time, important differences – with those obtained by Bhat [1] for the limiting distribution of the population in the finite buffer of a double-ended queue.

The distribution of time between kit completions, D_n, can be expressed in Laplace transform form as

$$L[\mathrm{d}P(\tau_n - \tau_{n-1} \leq t)] = \Pi Q(i, j, \mathrm{d}s)U, \tag{25}$$

in which U is a column vector with each element equal to 1. Substituting equations (21), (22), (23) and (24), equation (25) specializes to

$$L[\mathrm{d}P(\tau_n - \tau_{n-1} \leq t)] = \left(\frac{\mu}{\mu + s}\right)\left[1 - \left(\frac{2}{K_1 + K_2}\right)\left(\frac{s}{2\mu + s}\right)\right]. \tag{26}$$

Clearly, for large values of $K_1 + K_2$, the distribution of time between kit completions, D_n, is approximately exponential with rate μ. The value of $K_1 + K_2$ necessary to allow this approximation can be determined as a function of the degree of approximation desired. The ϵ-approximate distribution of D_n is

$$\Pi Q(i, j, \mathrm{d}s)U = L[\mathrm{d}P(\tau_n - \tau_{n-1} \leq t)] = \frac{\mu}{\mu + s}, \tag{27}$$

which is the Laplace transform of an exponential distribution with rate μ.

APPROXIMATE INDEPENDENCE OF D_n

In this section, we discuss the independence of m consecutive random variables D_n, $n = 1, 2, \ldots, m$. We show that for sufficiently large $K_1 + K_2$, the m consecutive random variables D_n, $n = 1, 2, \ldots, m$, become independent to within an error of ϵ.

This independence holds if the joint distribution of the m consecutive random variables D_n equals the product of the m marginal distributions of the random

variables D_n. Statistical independence should hold for $m \to \infty$, but this limiting case is not easily evaluated.

To establish the approximation, we must show (writing $Q(i, j, \mathrm{d}s) = Q(\mathrm{d}s)$),

$$\Pi Q(\mathrm{d}s_1)Q(\mathrm{d}s_2)Q(\mathrm{d}s_3)\ldots Q(\mathrm{d}s_m)U = \{\Pi Q(\mathrm{d}s_1)U\}\{\Pi Q(\mathrm{d}s_2)U\}\{\Pi Q(\mathrm{d}s_3)U\}\ldots\{\Pi Q(\mathrm{d}s_m)U\}. \tag{28}$$

The left hand side of equation (28) is

$$\begin{aligned}
&\Pi Q(\mathrm{d}s_1)Q(\mathrm{d}s_2)Q(\mathrm{d}s_3)\ldots Q(\mathrm{d}s_m)U \\
&= \prod_{i=1}^{m}\left(\frac{\mu}{\mu+s_i}\right)\left[1-\frac{2}{K_1+K_2}\right]+\frac{2}{K_1+K_2}\prod_{i=2}^{m}\left(\frac{\mu}{\mu+s_i}\right) \\
&\quad\times\left[b\left(\frac{1}{2}a\right)^{K_2}+\left(\frac{1}{2}a\right)^{K_2}+\ldots+\left(\frac{1}{2}a\right)^{3}+\left(\frac{1}{2}\right)a\prod_{i=2}^{m}\left(\frac{\mu}{\mu+s_i}\right)\right. \\
&\quad\left.+\left(\frac{1}{2}a\right)^{3}+\ldots+\left(\frac{1}{2}a\right)^{K_1}+b\left(\frac{1}{2}a\right)^{K_1}\right] \\
&= \prod_{i=1}^{m}\left(\frac{\mu}{\mu+s_i}\right)-\frac{2}{K_1+K_2}\left(\prod_{i=1}^{m}\left(\frac{\mu}{\mu+s_i}\right)-\prod_{i=2}^{m}\left(\frac{\mu}{\mu+s_i}\right)\right. \\
&\quad\times\left[b\left(\frac{1}{2}a\right)^{K_2}+\ldots+\left(\frac{1}{2}a\right)^{3}+\left(\frac{1}{2}\right)a\prod_{i=2}^{m}\left(\frac{\mu}{\mu+s_i}\right)\right. \\
&\quad\left.\left.+\left(\frac{1}{2}a\right)^{3}+\ldots+b\left(\frac{1}{2}a\right)^{K_1}\right]\right).
\end{aligned} \tag{29}$$

By making $K_1 + K_2$ sufficiently large, the right hand side of equation (29) can be approximated by $(\prod_{i=1}^{m}(\mu/(\mu+s_i))$, the product of the Laplace transform of the m marginal distributions of the random variables D_n for $n \in \mathbb{N}$. Hence, equation (28) holds for sufficiently large $K_1 + K_2$, indicating that the random variables D_n, $n = 1, \ldots, m$, are independent. The required value of $K_1 + K_2$ depends upon the degree of approximation desired.

The implications of equations (27) and (29) lead to the following theorem.

THEOREM 1

The arrival process of kits at buffer I_0 can be approximated by a Poisson process with rate μ, the degree of approximation depending on the value of $K_1 + K_2$. □

DEGREE OF APPROXIMATION: AN EXAMPLE

To illustrate the relationship between the degree of approximation of the arrival rate at I_0 and the buffer capacities K_1 (K_1), we consider the following example with equal buffer capacities $K_1 = K_2 = K$ and equal Poisson arrival rates $\mu_1 = \mu_2 = \mu$ at buffers I_1 and I_2, respectively.

Using equation (26), the density function of the time between kit completions, D_n, may be expressed in Laplace transform form as:

$$f(s) = \left(\frac{\mu}{\mu+s}\right) - \frac{1}{K}\left(\frac{\mu}{\mu+s}\right)\left(\frac{s}{2\mu+s}\right). \tag{30}$$

Inverting to the time domain, the density function of D_n is obtained as

$$f(t) = \mu e^{-\mu t} + \frac{\mu}{K} e^{-\mu t} - \frac{2\mu}{K} e^{-2\mu t}, \quad t \geq 0. \tag{31}$$

We define an error term $e(t)$, expressed as the absolute difference between the exponential density and the actual density of D_n:

$$e(t) = \frac{\mu}{K}|2e^{-2\mu t} - e^{-\mu t}|. \tag{32}$$

Using equation (31), graphs of $f(t)$ are plotted for $\mu = 1$ and $K = 2$, 5 and 10 against time $t \geq 0$ in figure 3. It is observed from figure 3 that the density of D_n rapidly approaches an exponential density as K increases. The graph of $e(t)$ against K, plotted for $\mu = 1$ and $t = 0.2$, is presented in figure 4, which also indicates that the error term $e(t)$ approaches zero rapidly as K increases.

Using equation (32), it is easily seen that for $\forall t, t > 0, |2e^{-2\mu t} - e^{-\mu t}| \leq 1$. Hence for a given $\epsilon > 0$ and for any arrival rate μ, we can find a K such that

$$\frac{\mu}{K} \leq \epsilon.$$

Therefore, the inventory capacity required to effect the desired approximation can be easily determined knowing the component arrival rate.

5. Discussion and conclusion

We have proven conditions for which the inter-arrival times of kits arriving to assembly are approximately independent and identically distributed exponential random variables. If components arrive at I_1 and I_2 according to independent and identical Poisson arrival streams and if $K_1 + K_2$ is sufficiently large, the output

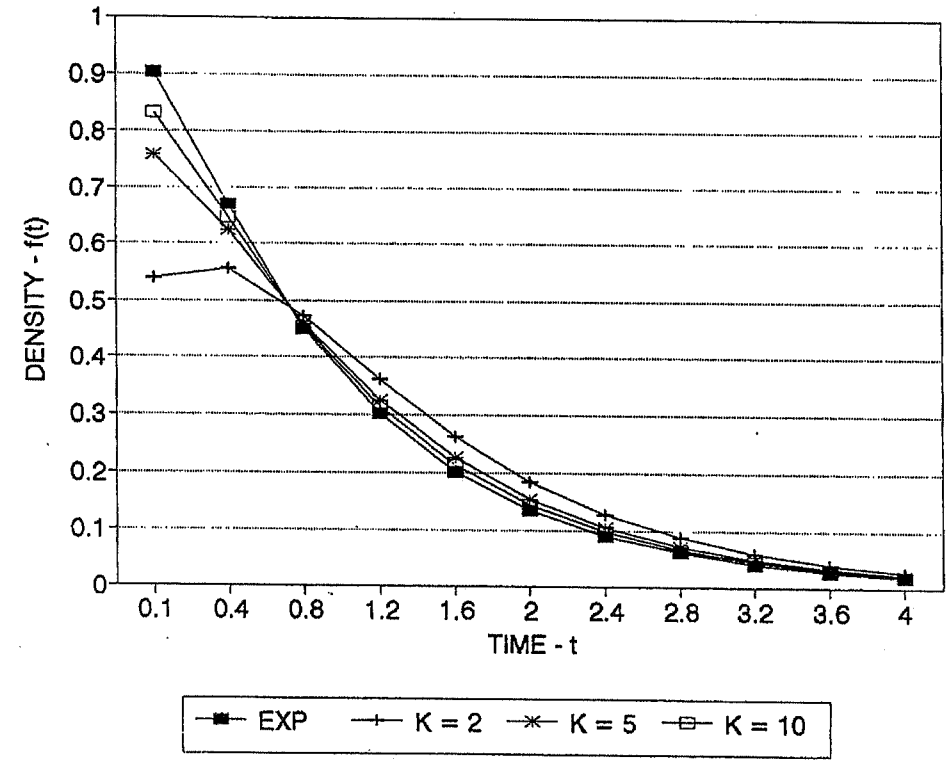

Fig. 3. Density comparison.

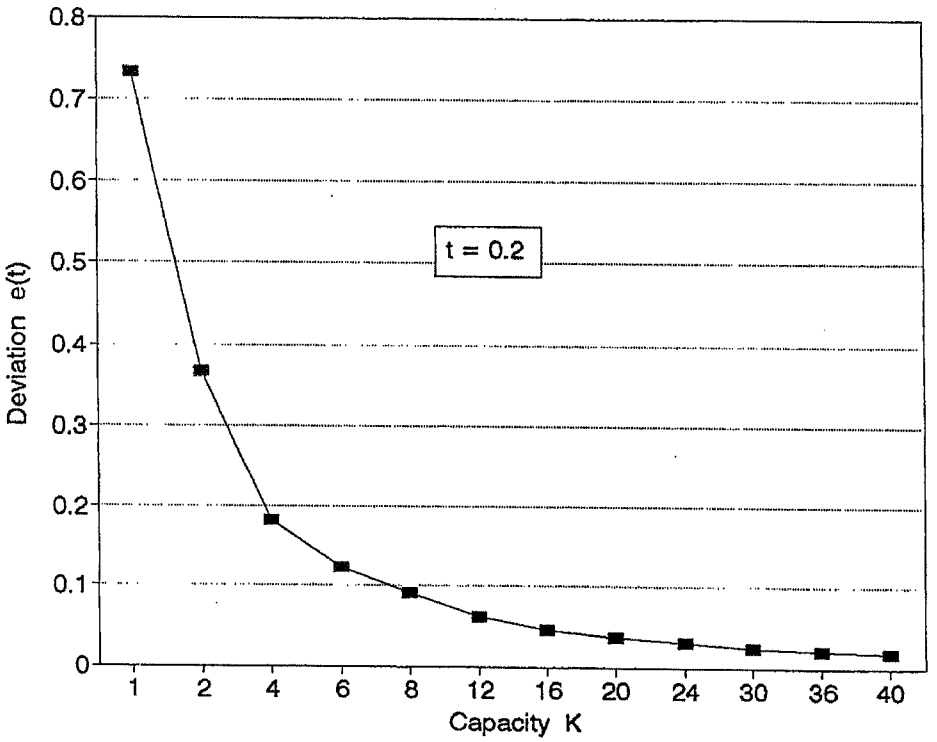

Fig. 4. Absolute deviation vs. bigger capacity.

stream from kitting approximates a Poisson process. The practical importance of this result is that the assembly process downstream of the kitting operation can be decoupled from kitting for further analysis. The required conditions (for decoupling) are not restrictive and may, in fact, hold in actual applications.

It is also interesting to note that the long-term probability distribution of inventory position j at I_1 (I_2), observed at kit completion epochs, depends on the inventory position j. If arrival rates to I_1 and I_2 are equal (i.e., $\mu_1 = \mu_2 = \mu$), all the inventory positions except zero become equally likely with probability that is inversely proportional to the total inventory capacity $(K_1 + K_2)$. The incidence of observing both buffers empty is twice as likely as observing a positive (negative) stock position at either of the buffers.

Harrison [6] showed that a sufficient condition for an assembly-like queue to reach steady state is that buffer capacities must be bounded from above. We have shown that the total buffer capacity, $K_1 + K_2$, must be "sufficiently large" to obtain a Poisson approximation of the output stream of kits. However, from the example in section 4, we find that $K_1 + K_2$ need not be impractically large to achieve an approximate Poisson output stream; the value of $K_1 + K_2$ being dependent upon the degree of approximation desired. Since the arrival process at assembly machine P_3 may be approximated by a Poisson distribution, the downstream assembly system can be approximated by the much studied $M/G/1$ queue.

Acknowledgements

This material is based on work supported by the National Science Foundation on Grant numbers DMC-8896303 and DDM-8913658. We are indebted to Dr. Hideo Ōsawa and two anonymous referees whose comments allowed us to strengthen an earlier version of this paper.

References

[1] U.N. Bhat, A controlled transportation queueing process, Manag. Sci. 16(1970)446–452.
[2] U.N. Bhat, Finite capacity assembly queues, Queueing Syst. 1(1986)85.
[3] R.L. Disney and P.C. Kiessler, *Traffic Processes Queuing Networks: A Markov Renewal Approach* (Johns Hopkins Univ. Press. 1987).
[4] R.L. Disney and D. Konig, Queueing networks: A survey of their random processes, SIAM Rev. 27 (1985) 335–403.
[5] J.M. Dobbie, A double-ended queueing problem of Kendall, Oper. Res. 9 (1961) 755–757.
[6] J.M. Harrison, Assembly-like queues, J. Appl. Prob. 10 (1973) 354–367.
[7] W.J. Hopp and J.T. Simon, Bounds and heuristics for assembly-like queues, Queueing Syst. 4 (1989) 137–156.
[8] B.R.K. Kashyap, A double-ended queueing system with limited waiting space, Proc. Natl. Inst. Sci. (India) 31 (1965) 559–570.
[9] B.R. Kashyap and M.L. Chaudhury, *An Introduction to Queuing Theory* (A&A Publ., Kingston, Ontario, Canada, 1988).
[10] G. Latouche, Queues with paired customers, J. Appl. Prob. 18 (1981) 684–696.

[11] S. Saboo and W.E. Wilhelm, An approach for modeling small-lot assembly networks, IIE Trans. (Dec. 1986) 322–334.
[12] P. Som and W.E. Wilhelm, Analysis of a stochastic assembly system – A Markov renewal approach, Working Paper, Texas A&M University (1992).
[13] W.E. Wilhelm, P. Som and B. Carroll, A model for implementing a paradigm of time-managed, material flow control in certain assembly systems, Int. J. Prod. Res. 30 (1992) 2063–2086.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers & Operations Research 32 (2005) 449–463
computers & operations research
www.elsevier.com/locate/dsw

Performance analysis of a kitting process in stochastic assembly systems

Satheesh Ramachandran[a], Dursun Delen[b,*]

[a]*Knowledge Based Systems, Inc., 1408 University Drive East, College Station, TX 77845, USA*
[b]*Department of MSIS, Oklahoma State University, 700 N. Greenwood Ave., Tulsa, OK 74106, USA*

Abstract

Kitting (accumulating components required for an assembly) plays a crucial role in determining the performance of a small-lot, multi-product, multi-level manufacturing system. In this paper, we analyze the kitting process as of a stochastic assembly system by treating it as an assembly-like queue. Specifically, we investigate the dynamics involved in a simple kitting process where two independent input streams feed into an assembly process. Unlike previous studies in this domain, we relax the assumption of finite buffer capacity constraint on the input buffers, and still show that the system remains stable under fairly mild conditions. It is expected that the findings of this study will provide manufacturing system designers with wider variety of control parameters to choose from in evaluating the system performance under a much broader set of control policies, which would lead to minimizing the associated costs.
© 2003 Elsevier Ltd. All rights reserved.

Keywords: Kitting; Assembly operations; Double-ended queue; Performance analysis

1. Introduction

Analysis of assembly operations plays a crucial role in improving the overall system performance in small-lot, multi-product, multi-level manufacturing operations, especially when the system operates under a stochastic environment [1,2]. According to Chen and Wilhelm [3] assembly operations form a significant portion of the overall product cycle time (hence the total manufacturing cost) in many industries including semiconductor manufacturing. Funk [4] reports that assembly operations comprise of up to 40% of total manufacturing cost in the electronics industry. Therefore, efficient control and management of assembly operations is crucial in reducing the cycle time of the final product.

* Corresponding author. Tel.: +1-918-594-8283; fax: +1-918-594-8281.
E-mail address: delen@okstate.edu (D. Delen).

0305-0548/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0305-0548(03)00247-8

Conventionally, analysis of assembly operations has been based on the assumption that the system operates deterministically. A more realistic analysis hinges on the recognition of the stochastic elements (i.e., random arrival and random service times) that influence the system. Component availability at the various buffers (and consequently, the delivery schedules) is significantly affected by these stochastic elements. The goal of this paper is to understand and evaluate the implications of kitting operations on the performance measures of assembly systems that operate under stochastic conditions.

Successful management of kitting operations increases the productivity of any assembly process [5]. In the electronics manufacturing industry, efficient kitting mechanisms simplify material flow and provide for better shop floor control [6]. Kitting operations are also studied at the level of production strategies such as MRP and JIT systems; where production is managed by either a push or a pull mechanism [7]. In such cases, efficient control of the kitting operations found to play an important role in lowering work in process (WIP) inventory and hence decreasing the operational cost. Researchers at the center for quick response manufacturing at the University of Wisconsin-Madison have been working on examining and comparing the analytical performance of push and pull production control strategies [8,9]. Their approximation models favors JIT (pull/Kanban) over MRP (push)-type production strategies.

Another domain that witnesses widespread use of kitting operations is the subcontracting practice in supply chain management [10], where subcontractors supply the individual components and services for the various products to the prime manufacturer and the manufacturer assembles the kits. One such application environment is found in the various US department-of-defense (DoD) aircraft repair depots (such as the Oklahoma City Air Logistics Center OC-ALC). In the shop floor lingo at the depots, a kit is an actual collection of parts needed to assemble an asset (such as a helicopter engine) to completion. Typically, these parts, which could either be manufactured internally or supplied by external contractors, are gathered in an assembly methodology to aid production. Given that kitting-type operations are commonly found in these environments, a central problem here is efficient control of the kit assembly process that optimizes the delivery of these kits based upon the actual upstream demand for these kits. One such recent initiative titled "lean value chain (LVC) for critical parts procurement" sponsored by the Air Force Wright Laboratory's Manufacturing Technology Directorate involved developing solutions that enable coordinated response to anticipated and known critical part problems [11,12]. A critical part is defined as any part whose anticipated or actual lack of availability will prevent on-time completion of the weapon system. Critical parts are often the result of ill-defined (or lack thereof) of control policies that dictate their delivery to the kitting process (historically, within the OC-ALC facilities such as the GE Rotor shop, there was little or no control policies for the parts ordering and procurement processes). The focus of the LVC project was to reengineer the processes of linking production with material/component procurement with the current effort being focused on developing and incorporating analytically driven control policies.

In a related ongoing research effort, Leung and Kamath [13] analyze a single-stage assembly system where two components are assembled into a single product via a kitting operation. Each component has its own finite buffer for temporary storage while waiting for its counterpart. When a pair of components is available, the components move into the assembly station, which has its own input buffer. They develop a model to approximately calculate performance measures, such as mean system time and mean queue length, when the component arrival

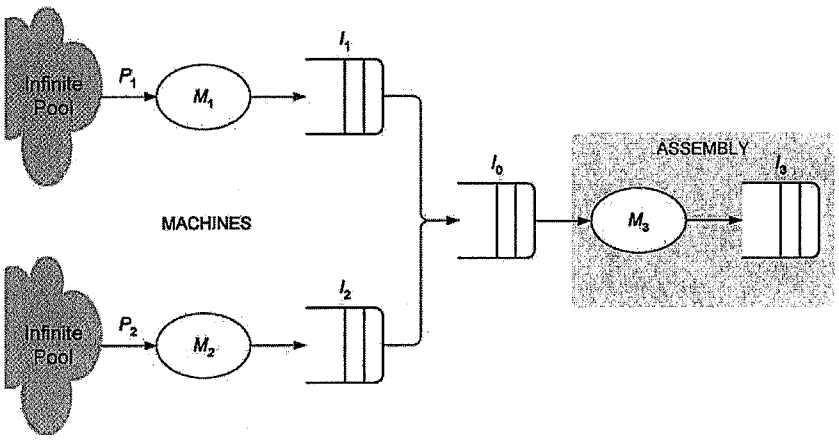

Fig. 1. The structure of the assembly system under investigation.

processes are Poisson and the assembly time follows a general distribution. In Kamath and Leung [14], this single-stage building block is used in the context of a production network to test the usefulness of the approximation developed. Chang and Chen [15] also looked at tandem queues as assembly-like queues in order to develop control policies that would increase the system performance measures.

In this paper, we investigate a simple kitting process with two input streams for an assembly system with the aim of understanding the dynamics involved. We assume that the arrival streams feeding the kitting process have state-dependent arrival intensities. The assembly system has a similar structure as modeled in [7,16], and is shown in Fig. 1.

In Fig. 1, M_1 and M_2 are machines processing parts P_1 and P_2, and M_3 assembles these machined components. I_1 and I_2 are the buffers for the machined components, I_0 is the buffer for the kitted component and I_3 is the buffer for the assembled component. Machines M_1 and M_2 are assumed to operate independently. They withdraw raw materials from their respective pools of infinite capacity and supply machined components to the buffers I_1 and I_2, respectively. A component arriving at buffer I_1 (I_2) is immediately kitted with a part from buffer I_2 (I_1), if one is available, and a kit is supposed to be ready for assembly operation at machine M_3. If the kit cannot be composed, the machined part is held in the buffer I_1 (I_2), and awaits the arrival of a "matching" part from I_2 (I_1). Once composed, the kit of matching components from I_1 and I_2 is sent to I_0 and the kit is considered to have arrived at I_0 (Fig. 2).

For exponential service times at M_1 and M_2 and finite buffer capacities at I_1 and I_2, Som et al. [16] characterize the occupancy distribution at I_1 and I_2 at kit departure epochs. Completed kits are shown to arrive at I_0 according to a Markov-renewal process. Also, when machines M_1 and M_2 have identical processing rates, and buffer capacities at I_1 and I_2 are large enough, they show that the arrival of completed kits to I_0 is well approximated by a Poisson process. This leads to the decoupling of the kitting operations from assembly, and hence to an easy analysis of the downstream assembly operations.

Stochastic assembly systems are often studied as assembly like queues [17,18]. Many followed the same approach in developing approximations for computing the performance measures of complex assembly operations [19,20]. Harrison [18] in a primarily theoretical study, established

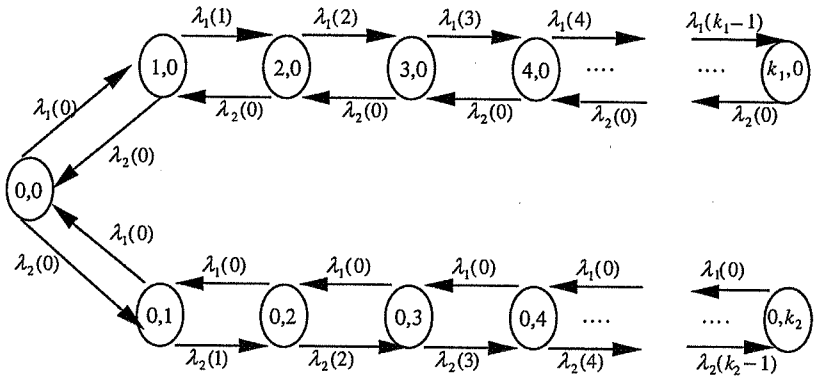

Fig. 2. Rate diagram for the kitting process.

stability conditions for an assembly queue with renewal and mutually independent arrival streams. He showed that a sufficient and necessary condition for the queue to be stable is for the component buffers at I_1 and I_2 to be finite. Thus, when there are no limitations on the inventory levels, the buffer sizes at I_1 and I_2 blow up, even when the arrival intensities (processing rates for M_1 and M_2) are the same. This can be intuitively explained by visualizing the queueing phenomena in the context of a double-ended queue [21]. A double-ended queue can best be described as the classical taxicab problem where taxies and passengers form two mutually separate queues [22]. A customer waits in the customer queue until a taxi is available, and taxies wait in the taxi queue until a customer is available. The two queues are interdependent and their combination is known as a double-ended queue. The underlying queueing process maps into a random walk on $\{\ldots,-2,-1,0,1,2,\ldots\}$, which is transient or recurrent null except when the queues are bounded. Hence, most of the analysis of assembly queues and kitting operations incorporates the finite buffer size assumption. A more realistic approach is to view the machine processing rates as control parameters, which dictate the performance measures of the system. The finite buffer capacities case is a specific policy for setting the control parameters which guarantees stability, but it need not be the optimal policy. The assumption of finite buffer capacity to ensure stability could be fairly restrictive. This is particularly true since in this case the system remains stable under fairly moderate conditions, allowing the system to be evaluated under a much broader set of control policies. This approach offers system designers with a wider variety of control parameters to choose from to minimize the associated costs. In this paper, we evaluate the system when the arrival rates (or the machine processing rates) to I_1 and I_2 are controlled as a function of the buffer sizes at I_1 and I_2, respectively. The service times of machines M_1 and M_2 are assumed to be exponential, and dependent on the buffer sizes at I_1 and I_2, respectively. Under these conditions in the next section, we characterize the probability laws for buffer sizes at I_1 and I_2 and establish conditions for system stability. Then we derive the waiting time distributions for kits arriving at I_0. We also show that waiting times degenerate to exponential waiting times under the conditions assumed by Som et al. [16].

The remainder of the paper is organized as follows. Section 2 presents the definitions, theorems and proofs of our approach. A simple numerical cost structure example is provided in Section 3. Section 4 concludes the paper by presenting the results and implications of our study.

2. Definitions and main results

Let $\lambda_1(n_1):n_1 \geqslant 0$ and $\lambda_2(n_2):n_2 \geqslant 0$ be the processing rate of machines M_1 and M_2, and respectively, where n_1 and n_2 are the number of machined components waiting at corresponding buffers I_1 and I_2. This is a generic characterization; for example, the special case for finite buffers of k_1 and k_2 at I_1 and I_2, respectively, and constant and identical machine processing times at M_1 and M_2 can be defined by the following conditions:

$$\lambda_1(n) = \lambda, \quad 0 \leqslant n \leqslant k_1$$
$$= 0, \quad n > k_1$$

$$\lambda_2(n) = \lambda, \quad 0 \leqslant n \leqslant k_2$$
$$= 0, \quad n > k_2.$$

In order to establish conditions on the control functions $\lambda_1(n_1):n_1 \geqslant 0$ and $\lambda_2(n_2):n_2 \geqslant 0$ which enable system stability, we characterize the probability laws of the number in system and establish the waiting time for the completed kits arriving at I_0 as a function of the processing rate at the machines M_1 and M_2. We show that when the inventory capacities at I_1 and I_2 are unlimited, the system is stable for very mild conditions on the control functions. Studying the behavior of the inventory levels at I_1 and I_2, and the departure rate as a function of the control parameters is useful in the selection of these parameters. We arrive at these stability conditions by first developing characterizations for a finite capacity system, and then developing the unlimited buffer size case as a limiting case of the finite capacity system.

Theorem 1. *Following the previous research* [2,6,16,17], *we let the service times of machines* M_1 *and* M_2 *be exponentially distributed with parameters* $\lambda_1(n_1)$, *and* $\lambda_2(n_2)$, *where* n_1 *and* n_2 *are the number of machined components at buffers* I_1 and I_2, *respectively. The permissible queue sizes in both stations are* k_1 *and* k_2, *respectively. Let* $\pi_{0,0}, \pi_{0,1}, \ldots, \pi_{0,k_2}, \pi_{1,0}, \ldots, \pi_{k_1,0}$ *be the steady-state probabilities for the system states and let*

$$L = \left\{ \begin{array}{l} 1 + \left[\dfrac{\lambda_1(0)}{\lambda_2(0)} + \dfrac{\lambda_1(0)\lambda_1(1)}{\lambda_2^2(0)} + \dfrac{\lambda_1(0)\lambda_1(1)\lambda_1(2)}{\lambda_2^3(0)} + \cdots + \dfrac{\lambda_1(0)\ldots\lambda_1(k_1-1)}{\lambda_2^{k_1}(0)} \right] \\ \quad + \left[\dfrac{\lambda_2(0)}{\lambda_1(0)} + \dfrac{\lambda_2(0)\lambda_2(1)}{\lambda_1^2(0)} + \dfrac{\lambda_2(0)\lambda_2(1)\lambda_2(2)}{\lambda_1^3(0)} + \cdots + \dfrac{\lambda_2(0)\ldots\lambda_2(k_2-1)}{\lambda_2^{k_2}(0)} \right] \end{array} \right\}.$$

Then,

$$\pi_{0,0} = \frac{1}{L},$$

$$\pi_{0,k_2} = \frac{\lambda_2(k_2-1)\ldots\lambda_2(0)}{\lambda_1^{k_2}(0)L},$$

$$\pi_{k_1,0} = \frac{\lambda_1(k_1-1)\ldots\lambda_1(0)}{\lambda_2^{k_1}(0)L},$$

$$\pi_{0,n} = \frac{\lambda_1^{k_2-n}(0)}{\lambda_2(k_2-1)\ldots\lambda_2(n)}\,\pi_{0,k_2} \quad \textit{for } 0 < n < k_2,$$

$$\pi_{n,0} = \frac{\lambda_2^{k_1-n}(0)}{\lambda_1(k_1-1)\ldots\lambda_1(n)}\,\pi_{k_1,0} \quad \textit{for } 0 < n < k_1.$$

Proof. Let the state space for this assembly process be described as a two-tuple (n_1, n_2), where n_1 and n_2 correspond to the number of parts in the buffers I_1 and I_2, respectively. The kitting process is such that if $n_1 > 0 (n_2 > 0)$ then $n_2 = 0 (n_1 = 0)$. Assuming an infinitesimal kitting time, (i.e., a part arriving at either buffer is immediately kitted with a part from the complementary buffer), we have if $n_1 > 0$, it follows that $n_2 = 0$ and vice versa.

The balance equation for state $(k_1, 0)$ gives

$$\pi_{k_1-1,0} = \frac{\lambda_2(0)}{\lambda_1(k_1-1)}\,\pi_{k_1,0}.$$

Similarly,

$$\pi_{k_1-2,0} = \frac{\lambda_2^2(0)}{\lambda_1(k_1-1)\lambda_1(k_1-2)}\,\pi_{k_1,0}$$

and in general for $0 < n < k_1$ it follows that

$$\pi_{n,0} = \frac{\lambda_2^{k_1-n}(0)}{\lambda_1(k_1-1)\ldots\lambda_1(n)}\,\pi_{k_1,0} \tag{1}$$

and

$$\pi_{0,0} = \frac{\lambda_2^{k_1}(0)}{\lambda_1(k_1-1)\ldots\lambda_1(0)}\,\pi_{k_1,0}. \tag{2}$$

By symmetry, we also have for $0 < n < k_2$

$$\pi_{0,n} = \frac{\lambda_1^{k_2-n}(0)}{\lambda_2(k_2-1)\ldots\lambda_2(n)}\,\pi_{0,k_2} \tag{3}$$

and

$$\pi_{0,0} = \frac{\lambda_1^{k_2}(0)}{\lambda_2(k_2-1)\ldots\lambda_2(0)}\,\pi_{0,k_2}.$$

Equating the two expressions for $\pi_{0,0}$, we get

$$\pi_{0,k_2} = \frac{\lambda_2^{k_1}(0)}{\lambda_1^{k_2}(0)}\,\frac{\lambda_2(k_1-1)\ldots\lambda_2(0)}{\lambda_1(k_1-1)\ldots\lambda_1(0)}\,\pi_{k_1,0}. \tag{4}$$

Substituting (3) in (2) we get

$$\pi_{0,n} = \frac{\lambda_2^{k_1}(0)}{\lambda_1^{n}(0)}\,\frac{\lambda_2(n-1)\ldots\lambda_2(0)}{\lambda_1(k_1-1)\ldots\lambda_1(0)}\,\pi_{k_1,0}. \tag{5}$$

The normalizing equation for this system is

$$(\pi_{k_1,0} + \pi_{k_1-1,0} + \pi_{k_1-2,0} + \pi_{k_1-3,0} + \cdots + \pi_{1,0}) + (\pi_{0,k_2} + \pi_{0,k_2-1} + \pi_{0,k_2-2} + \pi_{0,k_2-3} + \cdots + \pi_{0,1}) + \pi_{0,0} = 1.$$

Substituting for all the probabilities in terms of $\pi_{k_1,0}$, using Eqs. (1) and (5) in the above equation, we have

$$\pi_{k_1,0}\left[1 + \left\{\frac{\lambda_2(0)}{\lambda_1(k_1-1)} + \cdots + \frac{\lambda_2^{k_1-1}(0)}{\lambda_1(k_1-1)\ldots\lambda_1(1)} + \frac{\lambda_2^{k_1-1}(0)}{\lambda_1(k_1-1)\ldots\lambda_1(1)}\right\} + \left\{\frac{\lambda_2^{k_1}(0)}{\lambda_1^{k_2}(0)}\frac{\lambda_2(k_2-1)\ldots\lambda_2(0)}{\lambda_1(k_1-1)\ldots\lambda_1(0)} + \frac{\lambda_2^{k_1}(0)}{\lambda_1^{k_2-1}(0)}\frac{\lambda_2(k_2-2)\ldots\lambda_2(0)}{\lambda_1(k_1-1)\ldots\lambda_1(0)} + \cdots + \frac{\lambda_2^{k_1}(0)}{\lambda_1(0)}\frac{\lambda_2(0)}{\lambda_1(k_1-1)\ldots\lambda_1(0)}\right\}\right] = 1.$$

After employing Eq. (2), the expression for $\pi_{0,0}$ follows as $\pi_{0,0} = 1/L$, where L is as defined above. □

Next, we extend the finite capacity case to infinite buffers and derive some sufficient conditions for stability. For stability, we check conditions under which $\pi_{0,0} > 0$. This is equivalent to checking the condition that the series in the denominator of the expression for $\pi_{0,0}$ converges. The following theorem states the stability conditions for the control function of the kitting process.

Theorem 2. *If* $\lim_{k\to\infty}\lambda_1(k) < \lambda_2(0)$ *and* $\lim_{k\to\infty}\lambda_2(k) < \lambda_1(0)$ *then the system is stable.*

Proof. The queue is stable *iff* the series

$$1 + \left[\frac{\lambda_1(0)}{\lambda_2(0)} + \frac{\lambda_1(0)\lambda_1(1)}{\lambda_2^2(0)} + \frac{\lambda_1(0)\lambda_1(1)\lambda_1(2)}{\lambda_2^3(0)} + \cdots + \frac{\lambda_1(0)\ldots\lambda_1(k_1-1)}{\lambda_2^{k_1}(0)} + \cdots\right] + \left[\frac{\lambda_2(0)}{\lambda_1(0)} + \frac{\lambda_2(0)\lambda_2(1)}{\lambda_1^2(0)} + \frac{\lambda_2(0)\lambda_2(1)\lambda_2(2)}{\lambda_1^3(0)} + \cdots + \frac{\lambda_2(0)\ldots\lambda_2(k_2-1)}{\lambda_1^{k_2}(0)} + \cdots\right] \text{ converges.}$$

The series above has all positive terms. A sufficient condition for the above series to converge is that both of the following series converge.

$$1 + \frac{\lambda_1(0)}{\lambda_2(0)} + \frac{\lambda_1(0)\lambda_1(1)}{\lambda_2^2(0)} + \frac{\lambda_1(0)\lambda_1(1)\lambda_1(2)}{\lambda_2^3(0)} + \cdots,$$

and

$$1 + \frac{\lambda_2(0)}{\lambda_1(0)} + \frac{\lambda_2(0)\lambda_2(1)}{\lambda_1^2(0)} + \frac{\lambda_2(0)\lambda_2(1)\lambda_2(2)}{\lambda_1^3(0)} + \cdots.$$

Let the kth term in the first series be a_k and the kth term in the second series be b_k. Series 1 converges if, $\lim_{k\to\infty} a_{k+1}/a_k < 1$. Similarly, Series 2 converges if $\lim_{k\to\infty} b_{k+1}/b_k < 1$. We have,

$\lim_{k\to\infty} a_{k+1}/a_k = \lim_{k\to\infty} \lambda_1(k)/\lambda_2(0) < 1 \Rightarrow \lim_{k\to\infty} \lambda_1(k) < \lambda_2(0)$ and

$$\lim_{k\to\infty} \frac{b_{k+1}}{b_k} = \lim_{k\to\infty} \frac{\lambda_2(k)}{\lambda_1(0)} < 1 \Rightarrow \lim_{k\to\infty} \lambda_2(k) < \lambda_1(0),$$

thus proving the theorem. □

Based upon this result, it is evident that system stability is guaranteed under mild conditions on the control functions. Intuitively, the above result states that the system is guaranteed stability as long as the control functions $\lambda_2(k)$ and $\lambda_1(k)$ which represents the tendency to drift towards $(0,\infty)$ and $(\infty,0)$, respectively, are finally dominated by $\lambda_1(0)$ and $\lambda_2(0)$ (which represent the tendency of the system to pull back to the state $(0,0)$), respectively.

Next we establish the waiting time distribution for kits arriving at buffer I_0. Let $T_1, T_2, T_3, \ldots$ be the times of completion of successive kits. Let $X_1, X_2, X_3, \ldots$ be the queue sizes. Then Som et al. [16] show that when the maximum permissible buffer sizes k_1 and k_2 are finite (X_n, T_n) form a Markov renewal process. They develop expressions for $P\{T_{n+1} - T_n \leqslant t\}$ and show that it approximates an exponential distribution as k_1 and k_2 become infinitely large. We use the result in Theorem 1 under the more general assumptions of infinite buffer capacity and controlled arrival rates to characterize the waiting time distributions for kits arriving at I_0.

Theorem 3. *Let N_t be the number of kits completed up until time t. Let E be the state space for the process i.e., $E = \{(k_1,0),\ldots,(0,0),\ldots,(0,k_2)\}$. Let $T_{N_t+1} - t = W_{N_t+1}$. Then the distribution of the waiting time W_{N_t+1} at steady state is given by*

$$\lim_{t\to\infty} P\{W_{N_t+1} \leqslant y\} = A(1 - e^{-\lambda_2(0)y}) + B(1 - e^{-\lambda_1(0)y}) + (1 - A - B)(1 - e^{-\lambda_1(0)y})(1 - e^{-\lambda_2(0)y})$$

where

$$A = \sum_{n=1}^{\infty} \pi_{n,0} \quad \text{and} \quad B = \sum_{n=1}^{\infty} \pi_{0,n}.$$

Proof.

Case (a):

$$\lim_{t\to\infty} P\{W_{N_t+1} \leqslant y/\text{state of system} = (k,0),\ k > 0\}$$

$$= P\{\text{arrival at station } I_2 \text{ before time } y \text{ units/state of system} = (k,0),\ k > 0\}$$

$$= 1 - e^{-\lambda_2(0)y}.$$

Case (b):

$$\lim_{t\to\infty} P\{W_{N_t+1} \leqslant y/\text{state of system} = (0,k),\ k > 0\}$$

$$= P\{\text{arrival at station } I_1 \text{ before time } y \text{ units/state of system} = (0,k),\ k > 0\}$$

$$= 1 - e^{-\lambda_1(0)y}.$$

Case (c):

$$\lim_{t\to\infty} P\{W_{N_t+1} \leqslant y/\text{state of system} = (0,0)\}$$
$$= P\{\text{arrival at station } I_1 \,\&\, I_2 \text{ before time } y \text{ units/state of system} = (0,0)\}$$
$$= (1 - e^{-\lambda_1(0)y})(1 - e^{-\lambda_2(0)y}).$$

Let

$$A = \sum_{n=1}^{\infty} \pi_{n,0}, \quad \text{and} \quad B = \sum_{n=1}^{\infty} \pi_{0,n}.$$

Then, $\pi_{0,0} = 1 - A - B$, and we have,

$$\lim_{t\to\infty} P\{W_{N_t+1} \leqslant y\} = A(1 - e^{-\lambda_2(0)y}) + B(1 - e^{-\lambda_1(0)y}) + (1 - A - B)(1 - e^{-\lambda_1(0)y})(1 - e^{-\lambda_2(0)y}). \tag{6}$$

Next, we derive the joint distribution of two successive kit completions times from an arbitrary time t. □

Theorem 4. *Let $T_1, T_2, T_3, \ldots$ be the times of completion of successive kits. Let N_t be the number of kits completed up until time t. Let the buffer sizes be finite at k_1 and k_2, respectively. Let E be the state space for the process i.e., $E = \{(k_1,0),\ldots,(0,0),\ldots,(0,k_2)\}$. Then the joint distribution of the waiting time and the next inter-departure time is given by*

$$\lim_{t\to\infty} P\{T_{N_t+1} - t \leqslant y_1, T_{N_t+2} - T_{N_t+1} \leqslant y_2\}$$
$$= A'(1 - e^{-\lambda_2(0)y_1})(1 - e^{-\lambda_2(0)y_2}) + B'(1 - e^{-\lambda_1(0)y_1})(1 - e^{-\lambda_1(0)y_2})$$
$$+ \pi_{1,0}\left[\frac{\lambda_1(0)}{\lambda_1(0)+\lambda_2(0)}(1 - e^{-\lambda_2(0)y_1})(1 - e^{-\lambda_2(0)y_2}) + \frac{\lambda_2(0)}{\lambda_1(0)+\lambda_2(0)}(1 - e^{-\lambda_2(0)y_1})(1 - e^{-\lambda_2(0)y_2})(1 - e^{-\lambda_1(0)y_2})\right]$$
$$+ \pi_{0,1}\left[\frac{\lambda_1(0)}{\lambda_1(0)+\lambda_2(0)}(1 - e^{-\lambda_1(0)y_1})(1 - e^{-\lambda_1(0)y_2})(1 - e^{-\lambda_2(0)y_2}) + \frac{\lambda_2(0)}{\lambda_1(0)+\lambda_2(0)}(1 - e^{-\lambda_1(0)y_1})(1 - e^{-\lambda_1(0)y_2})\right]$$
$$+ \pi_{0,0}\left[\frac{\lambda_1(0)}{\lambda_1(0)+\lambda_2(0)}(1 - e^{-\lambda_1(0)y_1})(1 - e^{-\lambda_2(0)y_1})(1 - e^{-\lambda_2(0)y_2}) + \frac{\lambda_2(0)}{\lambda_1(0)+\lambda_2(0)}(1 - e^{-\lambda_1(0)y_1})(1 - e^{-\lambda_2(0)y_1})(1 - e^{-\lambda_1(0)y_2})\right], \tag{7}$$

where

$$A' = \sum_{n=1}^{k_1} \pi_{n,0} \quad \textit{and} \quad B' = \sum_{n=1}^{k_2} \pi_{0,n}.$$

Proof. We can write the above expression as

$$\begin{aligned} &P\{T_{N_t+1} - t \leqslant y_1, T_{N_t+2} - T_{N_t+1} \leqslant y_2\} \\ &= \sum_{i \in E} \pi_i P\{T_{N_t+1} - t \leqslant y_1, T_{N_t+2} - T_{N_t+1} \leqslant y_2 / \text{state at } t = i\} \\ &= \sum_{i \in E} \pi_i P\{T_{N_t+2} - T_{N_t+1} \leqslant y_2 / T_{N_t+1} - t \leqslant y_1\} P\{T_{N_t+1} - t \leqslant y_1 / \text{state at } t = i\}. \end{aligned} \tag{8}$$

Case (a) ($X_{N_t} = (k, 0),\ k > 1$):

$$P\{T_{N_t+2} - T_{N_t+1} \leqslant y_2 / T_{N_t+1} - t \leqslant y_1,\ \text{state at } t = i\} = (1 - \mathrm{e}^{-\lambda_2(0) y_2}). \tag{9}$$

Case (b) ($X_{N_t} = (0, k),\ k > 1$):

$$P\{T_{N_t+2} - T_{N_t+1} \leqslant y_2 / T_{N_t+1} - t \leqslant y_1,\ \text{state at } t = i\} = (1 - \mathrm{e}^{-\lambda_1(0) y_2}). \tag{10}$$

Case (c) ($X_{N_t} = (1, 0)$):

$$\begin{aligned} &P\{T_{N_t+2} - T_{N_t+1} \leqslant y_2 / T_{N_t+1} - t \leqslant y_1,\ \text{state at } t = i\} \\ &= \frac{\lambda_1(0)}{\lambda_1(0) + \lambda_2(0)} (1 - \mathrm{e}^{-\lambda_2(0) y_2}) + \frac{\lambda_2(0)}{\lambda_1(0) + \lambda_2(0)} (1 - \mathrm{e}^{-\lambda_1(0) y_2})(1 - \mathrm{e}^{-\lambda_2(0) y_2}). \end{aligned} \tag{11}$$

Case (d) ($X_{N_t} = (0, 1)$):

$$\begin{aligned} &P\{T_{N_t+2} - T_{N_t+1} \leqslant y_2 / T_{N_t+1} - t \leqslant y_1,\ \text{state at } t = i\} \\ &= \frac{\lambda_1(0)}{\lambda_1(0) + \lambda_2(0)} (1 - \mathrm{e}^{-\lambda_1(0) y_2})(1 - \mathrm{e}^{-\lambda_2(0) y_2}) + \frac{\lambda_2(0)}{\lambda_1(0) + \lambda_2(0)} (1 - \mathrm{e}^{-\lambda_1(0) y_2}). \end{aligned} \tag{12}$$

Case (e) ($X_{N_t} = (0, 0)$):

$$\begin{aligned} &P\{T_{N_t+2} - T_{N_t+1} \leqslant y_2 / T_{N_t+1} - t \leqslant y_1,\ \text{state at } t = i\} \\ &= \frac{\lambda_1(0)}{\lambda_1(0) + \lambda_2(0)} (1 - \mathrm{e}^{-\lambda_2(0) y_2}) + \frac{\lambda_2(0)}{\lambda_1(0) + \lambda_2(0)} (1 - \mathrm{e}^{-\lambda_1(0) y_2}). \end{aligned} \tag{13}$$

Substituting Eqs. (9) and (13) in Eq. (8) and using Theorem 3, we obtain the desired result.

Theorem 3 can be used to estimate expected remaining waiting times. Theorem 4 can be used to study the correlations structure of the kit completion process. Based upon Theorem 4, the following corollary states that the waiting times are independent under the assumption of $\lambda_1(n)=\lambda_2(n)=\lambda$. □

Corollary. *When both buffer capacities are infinite and $\lambda_1(n)=\lambda_2(n)=\lambda$ for all n, then the above joint distribution reduces to a product of two exponentially distributed intervals.*

Proof. Let $k_1=k_2=k$. Then, $\pi_i=1/(2k+1)\ \forall i\in E$. The right-hand side in Eq. (7) as given in Theorem 4, reduces to

$$\left(\frac{2k-2}{2k+1}+\frac{1}{2k+1}\right)(1-e^{-\lambda y_1})(1-e^{-\lambda y_2})$$
$$+\frac{1}{2k+1}[(1-e^{-\lambda y_1})(1-e^{-\lambda y_2})^2+(1-e^{-\lambda y_1})^2(1-e^{-\lambda y_2})]$$
$$+\frac{1}{2k+1}(1-e^{-\lambda y_1})(1-e^{-\lambda y_2})(2-e^{-\lambda y_1}-e^{-\lambda y_2}).$$

If we let $k\to\infty$, we have $\lim_{t\to\infty}P\{T_{N_t+2}-T_{N_t+1}\leqslant y_2, T_{N_t+1}-t\leqslant y_1\}=(1-e^{-\lambda y_1})(1-e^{-\lambda y_2})$. □

In the next section, we use simple numerical examples to gain further insights and select control parameters for minimizing overall system costs.

3. Numerical example

Based upon the results of the previous sections it is evident that system stability is guaranteed under reasonably mild conditions on the control functions. This poses the system designer with the following question: given a particular cost structure (such as the inventory holding cost, delivery rate requirements), from the class of control functions satisfying the stability criterion, what is the 'optimal' function for the processing rates? Consider the following set that defines sequence-tuples that satisfy the stability criterion:

$$S=\left\{(a_j, a_k): a_j=[a_{j,n}]_{n=0}^{\infty},\ a_k=[a_{k,n}]_{n=0}^{\infty} \text{ and } \lim_{n\to\infty} a_{j,n}<a_{k,0};\ \lim_{n\to\infty} a_{k,n}<a_{j,0}\right\}.$$

Any element of the set $(a_j, a_k)\in S$ is an admissible control policy. Then the overall cost function f can be generically defined in terms of the control policy and the cost parameters as

$$\textit{cost function}=f((a_j, a_k),\ \textit{cost parameters}),$$

and the optimal control policy (a_j^*, a_k^*) satisfies

$$f(a_j^*, a_k^*)=\min_{All(a_j, a_k)\in S} f((a_j, a_k),\ \textit{cost parameters}).$$

Intuitively it can be reasoned that there exists no particular control function that minimizes total cost over all cost structures. In other words, the nature of the control function would be dependent on particular cost structure that is present in the application domain. Also, the set of control policies that are admissible in a domain would be dictated by the capacities of the machines producing

the individual components (machines M_1 and M_2). Then the set of admissible control policies are restricted to

$$S = \left\{ \begin{array}{l} (a_j, a_k): a_j = [a_{j,n}]_{n=0}^{\infty},\ a_k = [a_{k,n}]_{n=0}^{\infty} \text{ and } \lim_{n\to\infty} a_{j,n} < a_{k,0};\ \lim_{n\to\infty} a_{k,n} < a_{j,0}; \\ \qquad\qquad \sup[a_{j,n}]_{n=0}^{\infty} \leqslant \lambda_{\max}^1;\ \sup[a_{k,n}]_{n=0}^{\infty} \leqslant \lambda_{\max}^2 \end{array} \right\},$$

where $\lambda_{\max}^1$ and $\lambda_{\max}^2$ are the upper limits for production capacities at machines M_1 and M_2, respectively. A closed-form analytical solution to the problem defined above is difficult, and is the focus of our ongoing research investigation. Nevertheless, we can leverage the research results from the previous sections to develop simple, yet practical, control strategies (which although not rigorous, they offer some level of control on the operational costs). We show a sample numerical exercise that illustrates the application of the theoretical results. Consider three specific classes of control functions for processing rates at the machines that satisfy the stability criterion.

$$(1)\ a_{j,n} = a_{k,n} = \lambda_1(n) = \lambda_2(n) = C_0\left(\frac{1}{r^n}\right); \quad r > 1,\ n \geqslant 0,$$

$$(2)\ a_{j,n} = a_{k,n} = \lambda_1(n) = \lambda_2(n) = \begin{cases} C_1\left(\dfrac{1}{n}\right), & n > 0, \\ C_2, & n = 0. \end{cases}$$

$$(3)\ a_{j,n} = a_{k,n} = \lambda_1(n) = \lambda_2(n) = \begin{cases} C_3\left(1 - \dfrac{n}{M}\right), & 0 \leqslant n \leqslant M, \\ 0 & \text{otherwise.} \end{cases}$$

All of these control functions belong to a specific control function type that can be tuned in terms of parameters (we can call this the *control function parameter*). For example, the first control function type is a geometric control function, in which by modifying the parameter r, different production intensities (and different cost performances) can be achieved. A simple formulation for the cost function would incorporates a tradeoff between the *lateness cost* of assembled kits (cost that is proportional to the waiting time for completed kits) and the *holding cost* of components at the individual buffers. The lateness cost reduces when the inter-departure times of successive assemblies have a lower mean (kits are generated at a higher rate). If we tend to have a high number of parts in the buffers (higher *holding costs*), then we tend to move away from the situation in which both buffers are starved, which reduces the mean of the inter-departure times. Inherent in the notion of a *lateness cost* is the assumption of an infinite demand of assembled kits at the downstream buffer I_0. Let

C_h = Holding cost of a part in buffer I_1 or I_2.

C_l = Lateness cost of an assembled part.

I = Total number of parts in buffers I_1 and I_2.

W = Remaining waiting time of an assembled kit.

TC = Total cost.

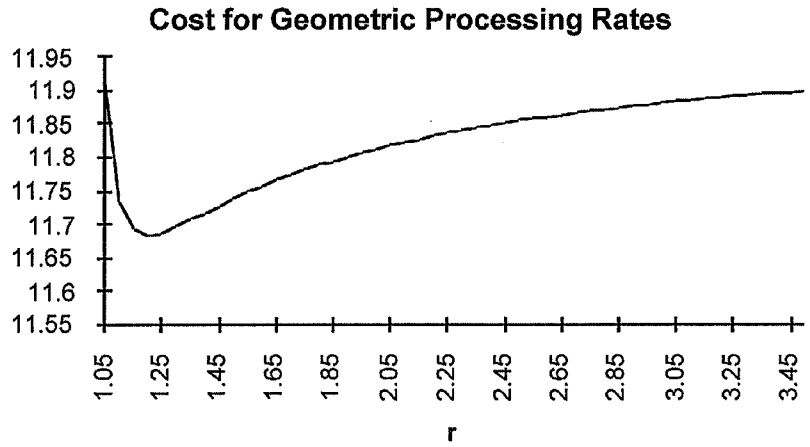

Fig. 3. Cost values for geometric processing rates with $C_h = 0.5$ and $C_l = 5$.

Table 1
Total cost at different cost combinations for three machine processing rates

Control policies	Policy 1, $C_h = 0.5, C_l = 5$, $C_0 = 1$, optimal r	Policy 2, $C_h = 0.5, C_l = 5$, $C_1 = 1, C_2 = 1$	Policy 3, $C_h = 0.5, C_l = 5$, $C_3 = 1, M = 10,000$
Cost	11.683	11.621	12.863

Then, define the total system costs as

$$TC = C_h E[I] + C_l E[W],$$

$E[I]$ is the expected number of components at buffers I_1 and I_2, and is defined as

$$E[I] = \sum_{i=0}^{\infty} iP(I = i).$$

The results from Theorem 1 can be used to compute $P(I = i)$. Similarly, if $f(y)$ is the density corresponding to the distribution in Theorem 3, then the expected remaining waiting time $E[W]$ of an assembled kit is defined as follows:

$$E[W] = \int_0^{\infty} yf(y)\,dy.$$

For a particular combination of $C_h (= 0.5)$ and $C_l (= 5)$, and parameter $C_0 = 1$, Fig. 3 shows the values for TC as a function of r (achieving an optimal cost value of 11.683). The entries in Table 1 compare this optimal TC value with the second (for $C_1 = 1, C_2 = 1$) and the third (for $C_3 = 1, M = 10,000$) control policies. We can see that the second processing rate (where the processing rate equals the reciprocal of the number of parts in the buffer) performs the best.

Although the above analysis makes simplifying assumptions and employs a primitive cost structure, it serves to illustrate the relevance of using machine processing rate functions as control parameters in improving the performance of assembly operations.

4. Conclusions

In this paper, we characterize probability laws for queue sizes at buffers for a kitting process. We derive the distribution of remaining waiting time for kits feeding the downstream process. We show that queue sizes at the component buffers are stable under very mild conditions for the control functions. This is in contrast to previous research, which analyses kitting systems of finite component buffer capacities, where the finite buffer sizes are imposed to ensure stability. This offers system designers a wide variety of control functions to choose from so as to have the flexibility to minimize cost given the cost structure at hand.

Acknowledgements

The authors would like to express their gratitude to Professor W.E. Wilhelm and to Professor Manjunath Kamath for their invaluable comments and directions for improving the content and presentation of this paper.

References

[1] Takahashi M, Osawa H, Fujisawa T. A stochastic assembly system with resume levels. Asia-Pacific Journal of Operational Research 1998;15:127–46.
[2] Chen JF, Wilhelm WE. An evaluation of heuristics for allocating components to kits in small-lot, multi-echelon assembly systems. International Journal of Production Research 1993;31(12):2835–56.
[3] Chen JF, Wilhelm WE. Kitting in multi-echelon, multi-product assembly systems with part substitutable. International Journal of Production Research 1997;35(10):2871–97.
[4] Funk JL. The potential market for robot assembly. International Journal of Production Research 1986;24(3):663–86.
[5] Ding FY. Kitting in JIT production: a kitting project at a tractor plant. IE Solutions, September 1992. p. 42–4.
[6] Wilhelm WE, Som P, Carroll B. A model for implementing a paradigm of time managed material flow control in certain assembly systems. International Journal of Production Research 1992;30(9):2063–86.
[7] Wilhelm WE, Som P. Analysis of single-stage, single-product, stochastic, MRP controlled assembly system. European Journal of Operations Research 1998;108:74–93.
[8] Ananth K, Suri R, Vernon M. Re-examining the performance of push, pull and hybrid material control strategies for multi-product flexible manufacturing systems. Technical Report. December 2000. p. 1–28.
[9] Krishnamurthy A, Suri R, Vernon M. A new approach for analyzing queueing models of material control strategies in manufacturing systems. In: Proceedings of the Fourth International Workshop on Queueing Networks with Finite Capacity (QNETs2000), West Yorkshire, U.K., July 2000.
[10] Simchi-Levi D, Kaminsky P, Simchi-Levi E. Designing and managing the supply chain, 2nd ed. New York: McGraw-Hill/Irwin; 2003.
[11] Painter M, Mayer R. Lean value chain for critical parts procurement. Final report prepared for air force wright laboratory's materials and manufacturing technology directorate—contract number F33615-98-C-5168, 2002.
[12] Benjamin P, Delen D, Lo A, Painter M, Graul M. A critical parts dashboard (CPD). Proceedings of the International Conference on Artificial Intelligence (IC-AI'2001), Las Vegas, Nevada: CSREA Press, June 25–28, 2001. p. 1562–7.
[13] Leung YT, Kamath M. Performance analysis of single-stage assembly system. ORSA/TIMS Joint National Meeting, 1994.
[14] Kamath M, Leung YT. Performance analysis of production networks involving assembly operations. INFORMS Los Angeles National Meeting, Spring 1995.
[15] Chang KH, Chen WF. Admission control policies for two-stage tandem queues with no waiting spaces. Computers and Operations Research 2003;30:589–601.

[16] Som P, Wilhelm WE, Disney RL. Kitting process in a stochastic assembly system. Queueing Systems 1994;17: 471–90.
[17] Hopp WJ, Simon JT. Bounds and Heuristics for assembly-like queues. Queueing Systems: Theory and Applications 1989;4:137–56.
[18] Harrison JM. Assembly-like queues. Journal of Applied Probability 1971;10:354–67.
[19] Hemachandra N, Eedupuganti SK. Performance analysis and buffer allocations in some open assembly systems. Computers and Operations Research 2003;30(5):695–704.
[20] Zhuang L, Wong YS, Fuh JYH, Yee CY. On the role of a queueing network model in the design of a complex assembly system. Robotics and Computer-Integrated Manufacturing 1998;14:153–61.
[21] Dobbie JM. A double-ended queueing problem of Kendall. Operations Research 1961;9:755–7.
[22] Kashyap BRK. A double ended queueing system with limited waiting space. Proceedings of the National Institute of Science of India 1965;31:559–70.

Queueing Systems, 4 (1989) 137–156

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

© J.C. Baltzer A.G. Scientific Publishing Company

[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

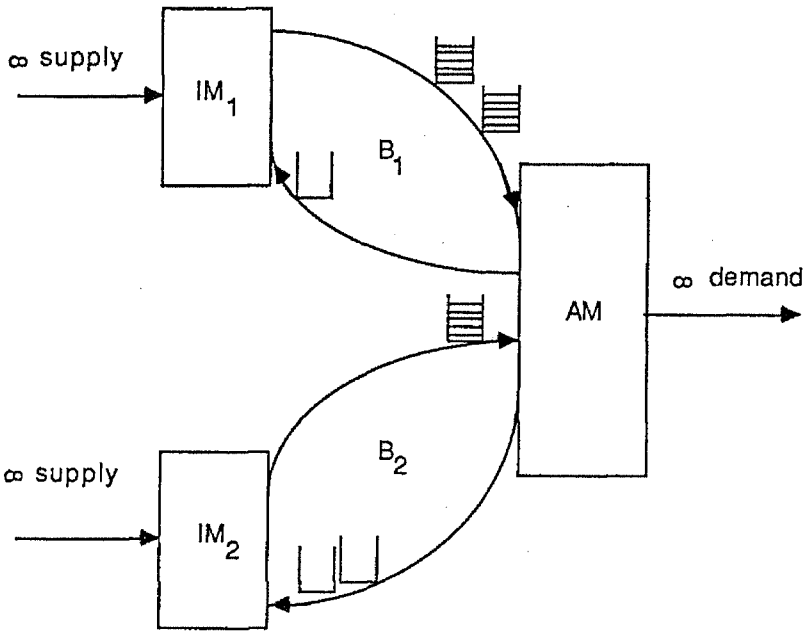

Fig. 1.

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. $\theta\{\lambda/\mu/1/K\}$, $\mathscr{I}\{\lambda/\mu/1/K\}$ 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 $\theta\{\lambda_1/\lambda_2/\mu/I\mathrm{K}_1/K_2\}$, $\mathscr{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ and $\mathscr{I}_2\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ 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, $\mathscr{I}_1$ and $\mathscr{I}_2$).

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

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

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

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

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'

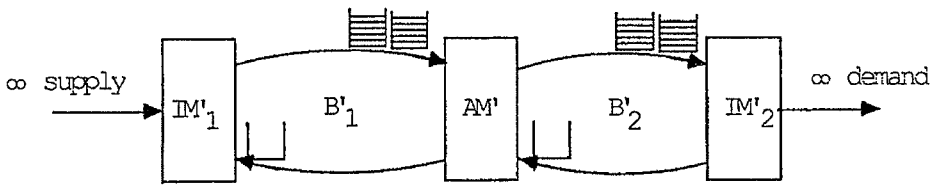

Fig. 2.

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].

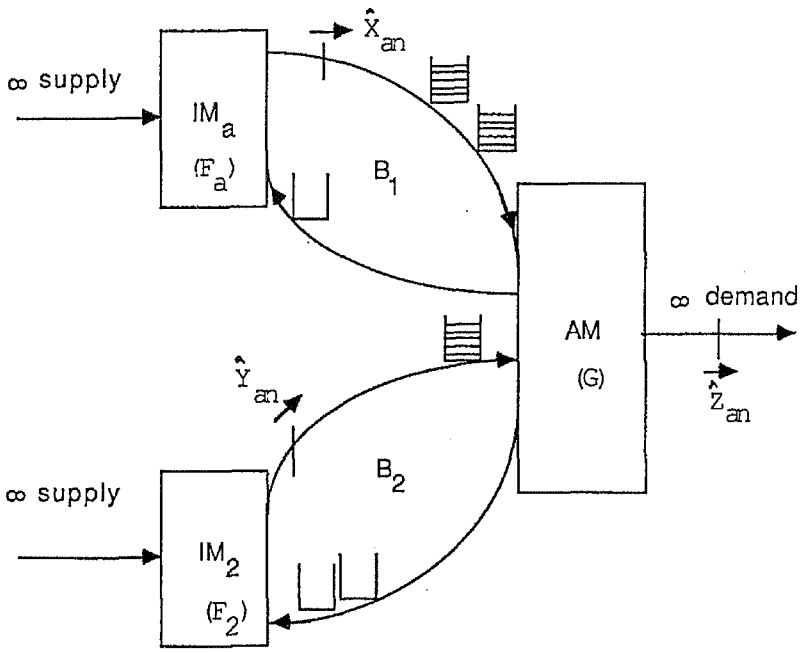

Fig. 3.

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

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

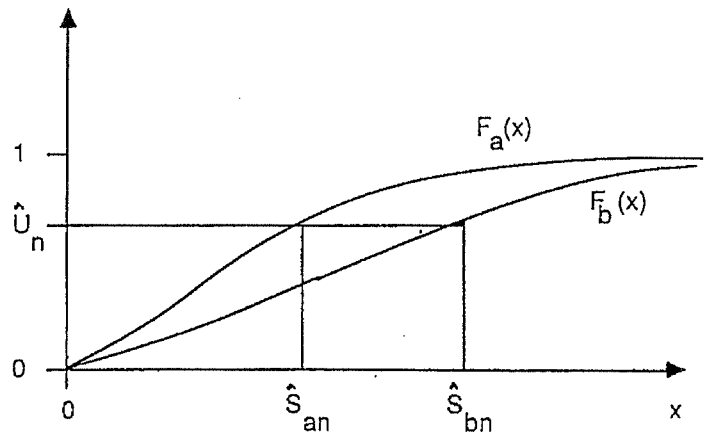

Fig. 4.

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

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

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

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

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

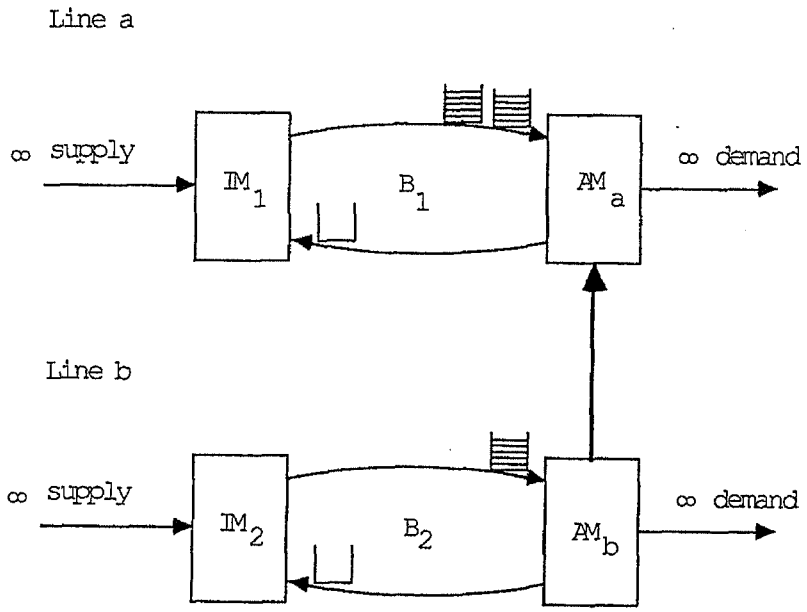

Fig. 5.

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:

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

As $\lambda_1 \to \infty$, $\theta_{hlb} \uparrow \theta\{\lambda_2/\mu/1/K_2\}$. Likewise, as $\lambda_2 \to \infty$, $\theta_{hlb} \uparrow \theta\{\lambda_1/\mu/1/K_1\}$. 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 $\mu \to \epsilon$, $\theta_{hlb} \uparrow \max\{\theta\{\lambda_1/\lambda_2/1/K_1\}, \theta\{\lambda_2/\lambda_1/1/K_2\}\}$. 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 $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$,

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

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

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

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

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 $\theta\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$,

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

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

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,

$$\mathcal{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} = \sum_{n=1}^{K_1} nP(N_1 = n)$$
$$\geq K_1 P(N_1 = K_1)$$
$$\geq K_1(1 - \theta_{ub}/\lambda_1).$$

The second inequality follows analogously. □

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

LEMMA 9

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

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 $\mathcal{I}_{lb1c}$ and $\mathcal{I}_{lb2c}$. Then $\mathcal{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geq \mathcal{I}_{lb1c}$ and $\mathcal{I}_2\{\lambda_1/\lambda_2/\mu/K_1/K_2\} \geq \mathcal{I}_{lb2c}$.

Proof

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

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

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

THEOREM 4

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

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:

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

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 $\mathcal{I}_{\text{hub1}} \leqslant \mathcal{I}_1\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$ or $\mathcal{I}_{\text{hub2}} \leqslant \mathcal{I}_2\{\lambda_1/\lambda_2/\mu/K_1/K_2\}$.

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).