

Lean Buffering in Serial Production Lines With Nonidentical Exponential Machines

Shu-Yin Chiang, Alexander Hu, and Semyon M. Meerkov, *Fellow, IEEE*

Abstract—Lean buffering is the smallest buffer capacity, which is necessary and sufficient to ensure the desired production rate of a manufacturing system. Literature offers methods for designing lean buffering in production systems with identical machines. The current paper extends these methods to serial production lines with nonidentical machines, assuming that they obey the exponential reliability model. For two-machine lines, exact formulas for lean buffering are derived, while for longer lines estimates are obtained. These results can be useful for production line designers and production managers to maintain the required production rate with the smallest possible inventories.

Note to Practitioners—In production systems with unreliable machines, operation with no-buffers (i.e., JIT) leads to low throughput. Very large buffers lead to high throughput but undesirable quality and economics properties. So, which level of buffering is good? This is the question addressed in this paper. The good level of buffering is addressed in terms of line efficiency, i.e., the fraction of the maximum of the throughput, which is acceptable for the system. For example, assume that the desired line efficiency is 0.9 (i.e., 90% of the maximum throughput is viewed as satisfactory). Under such an assumption, this paper offers methods for calculating the smallest (i.e., lean) level of buffering, which guarantees the desired throughput, provided that uptime and downtime of the machines are distributed exponentially with arbitrary mean time to failure and mean time to repair.

Index Terms—Exponential machines, lean buffering, serial production lines.

I. INTRODUCTION

LEAN buffering is the smallest buffer capacity, which is necessary and sufficient to achieve the desired production rate of a manufacturing system. Analytical methods for designing lean buffering in serial production lines with M identical exponential machines have been developed in [1]. The term “identical exponential” has been used to indicate that each machine has its uptime and downtime distributed exponentially with parameters p and r , respectively. In practice, however, this is seldom the case: even when the machines are exponential, each of them typically has different p_i and r_i , $i = 1, \dots, M$. This is referred to as the case of nonidentical machines. Methods for designing lean buffering in such systems is the

Manuscript received September 17, 2006. This paper was recommended for publication by Associate Editor S. Viswanathan and Editor N. Viswanadham upon evaluation of the reviewers’ comments. This work was supported in part by the National Science Foundation (NSF) under Grant DMI 0245377.

S.-Y. Chiang is with the Department of Information and Telecommunications Engineering, Ming Chuan University, Taoyuan 333, Taiwan, R.O.C.

A. Hu and S. M. Meerkov are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122 USA (e-mail: smm@eecs.umich.edu).

Digital Object Identifier 10.1109/TASE.2007.893503

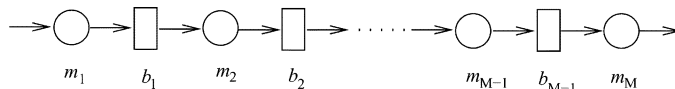


Fig. 1. Serial production line.

topic of this paper (see Section II for a precise model and problem formulation).

Specifically, we derive closed-form expressions for lean buffering in two-machine lines, investigate six approximation techniques (four based on closed formulas and two on recursions) for longer lines, and determine which one of them is most preferable under various conditions.

To place the current paper in the framework of existing literature, we remark (see [1] for a more detailed review) that the problem of buffer capacity allocation in serial lines has been studied quantitatively for over 50 years, and a large number of publications is available. The part of the literature, which is the closest to the current work, seeks the smallest total buffer capacity and its allocation so that the desired throughput is achieved. Both algorithmic and rule-based approaches to this problem have been investigated in [2]–[5] and [6]–[8], respectively. However, a rule-based approach to production lines with nonidentical exponential machines has not been developed. This is carried out in this paper.

The outline of this paper is as follows. Section II presents the model considered and the problem formulation. In Section III, the approach is described. Sections IV and V present the main results—for two and $M > 2$ nonidentical machines, respectively. Finally, in Section VI, the conclusions are formulated.

II. PROBLEM FORMULATION

A. Model

Consider a serial production line shown in Fig. 1 where the circles represent the machines and the rectangles are the buffers. Assume the line operates according to the following assumptions.

- i) The system consists of M machines arranged serially, and $M - 1$ buffers separating each consecutive pair of machines.
- ii) Each machine m_i has two states: up and down. When up, the machine is capable of producing one part per time interval referred to as the cycle time; when down, no production takes place; cycle times of all machines are assumed to be the same.
- iii) The uptime and downtime of each machine m_i are random variables distributed exponentially with parameters p_i and r_i , $i = 1, \dots, M$, respectively.

- iv) Each buffer is characterized by its capacity, $0 < N_i < \infty$, $i = 1, \dots, M - 1$.
- v) Machine m_i is starved at time t if buffer b_{i-1} is empty at time t . Machine m_1 is never starved.
- vi) Machine m_i is blocked at time t if buffer b_i has N_i parts at time t and machine m_{i+1} fails to take any work from this buffer at time t . Machine m_M is never blocked.

Assumption (iii) implies that the machine efficiency is

$$e_i = \frac{r_i}{r_i + p_i} = \frac{T_{\text{up},i}}{T_{\text{up},i} + T_{\text{down},i}}, \quad i = 1, \dots, M \quad (1)$$

where $T_{\text{up},i} = 1/p_i$ and $T_{\text{down},i} = 1/r_i$ are the average up-time and downtime of machine m_i , respectively. Assumptions (v) and (vi) indicate that the flow model of line operation is employed.

Model (i)–(vi) differs from the model analyzed in [1] in that [1] considered all identical machines and buffers, i.e.,

$$p_i = p, \quad r_i = r, \quad N_i = N \quad \forall i.$$

That is why the current model is referred to as a model with nonidentical machines. In both cases, however, the cycle times of all machines are the same.

B. Parametrization

Let PR denote the production rate of the line, i.e., the average number of parts produced by the last machine in the steady state. Let PR_∞ be the production rate of the line consisting of the same machines but having buffers with infinite capacity (and, therefore, having the largest possible production rate). Following [1], introduce the notion of *line efficiency*:

$$E = \frac{PR}{PR_\infty}. \quad (2)$$

Given the production line defined by assumptions (i)–(vi) and line efficiency E , the *lean buffer capacity* (LBC) is the sequence

$$N_{1,E}, \dots, N_{M-1,E} \quad (3)$$

such the desired line efficiency E is achieved, while $\sum_{i=1}^{M-1} N_{i,E}$ is minimized.

The LBC does not explicitly characterize whether the buffers are large or small. Indeed, a given LBC can be viewed as large if $T_{\text{down},i}$ are small and as small if $T_{\text{down},i}$ are large. Therefore, it

is convenient to represent LBC in units of the average downtime. This representation is referred to as the *lean level of buffering* (LLB) and is defined by

$$k_{i,E} = \frac{N_{i,E}}{\max\{T_{\text{down},1}, \dots, T_{\text{down},M}\}}, \quad i = 1, \dots, M - 1. \quad (4)$$

C. Problem

The problem considered in this work is as follows: Given the production line defined by assumptions (i)–(vi) and the desired line efficiency E , *develop analytical methods for calculating LBC*

$$N_{1,E}, \dots, N_{M-1,E}$$

and LLB

$$k_{1,E}, \dots, k_{M-1,E} \quad (5)$$

as functions of machine up parameter p_i , $i = 1, \dots, M$, machine down parameter r_i , $i = 1, \dots, M$, and the number of machines in the system M .

For the case of two-machine lines, this problem is solved in Section IV, longer lines are discussed in Section V, while the background material is described in Section III.

III. APPROACH

The approach of this paper is based on the methods for performance analysis and bottleneck identification developed in [9]–[12]. In addition, one of the recursive techniques uses the method for LLB evaluation in serial production lines with identical machines derived in [1]. To make this paper self-contained, these techniques are briefly reviewed below.

A. Production Rate Evaluation

A method for calculating PR of a line defined by assumptions (i)–(vi) with $M = 2$ has been developed in [9]. Specifically, it has been shown that

$$\begin{aligned} PR &= e_1[1 - Q(p_2, r_2, p_1, r_1, N)] \\ &= e_2[1 - Q(p_1, r_1, p_2, r_2, N)] \end{aligned} \quad (6)$$

where e_i , $i = 1, 2$, is the machine efficiency

$$e_i = \frac{r_i}{r_i + p_i} \quad (7)$$

and (8) and (9) at the bottom of the page.

$$Q(p_1, r_1, p_2, r_2, N_1) = \begin{cases} \frac{(1-e_1)(1-\phi)}{1-\phi \exp\{-\beta N_1\}}, & \text{if } \frac{p_1}{r_1} \neq \frac{p_2}{r_2} \\ \frac{p_1(p_1+p_2)(r_1+r_2)}{(p_1+r_1)[(p_1+p_2)(r_1+r_2)+p_2r_1(p_1+p_2+r_1+r_2)N_1]}, & \text{if } \frac{p_1}{r_1} = \frac{p_2}{r_2} \end{cases} \quad (8)$$

$$\begin{aligned} \phi &= \frac{e_1(1-e_2)}{e_2(1-e_1)} \\ \beta &= \frac{(r_1+r_2+p_1+p_2)(p_1r_2-p_2r_1)}{(r_1+r_2)(p_1+p_2)} \end{aligned} \quad (9)$$

It has been shown in [9] that the meaning of function Q is that $e_1Q(p_2, r_2, p_1, r_1, N)$ is the probability of the first machine to be blocked and $e_2Q(p_1, r_1, p_2, r_2, N)$ is the probability of starvation of the second machine.

For $M > 2$, no explicit formulas for PR have been derived and, therefore, recursive aggregation procedures, which lead to an estimate of PR with the accuracy typically within 1%, have been developed (see [10] and [11]). The procedures consist of two parts—the backward and forward aggregations. Within the backward aggregation, the last two machines are aggregated into a single machine, using expressions (6)–(9). Then, this aggregated machine is aggregated with the $(M-2)$ th machine and so on, until all machines are aggregated into a single one. The uptime and downtime parameters of the aggregated machines are denoted as p_i^b and r_i^b , $i = 1, \dots, M-1$, respectively, where b stands for “backward.” In the forward aggregation, the first machine is aggregated with the aggregated machine representing the last $M-1$ machines. Then, this machine is aggregated with the aggregation of the last $M-2$ machines and so on, until all machines are again aggregated into one. The uptime and downtime parameters of these aggregated machines are denoted as p_i^f and r_i^f , $i = 2, \dots, M$, respectively, where f stands for “forward.” Then, the procedure is repeated again, alternating between the backward and forward aggregations. Formally, this process is represented as (see [11])

$$\begin{aligned}
 r_i^b(s+1) &= r_i - r_i Q\left(p_{i+1}^b(s+1), r_{i+1}^b(s+1), \right. \\
 &\quad \left. p_i^f(s), r_i^f(s), N_i\right), \\
 &\quad i = 1, \dots, M-1, \\
 p_i^b(s+1) &= p_i + r_i Q\left(p_{i+1}^b(s+1), r_{i+1}^b(s+1), \right. \\
 &\quad \left. p_i^f(s), r_i^f(s), N_i\right), \\
 &\quad i = 1, \dots, M-1, \\
 r_i^f(s+1) &= r_i - r_i Q\left(p_{i-1}^f(s+1), r_{i-1}^f(s+1), \right. \\
 &\quad \left. p_i^b(s+1), r_i^b(s+1), N_{i-1}\right), \\
 &\quad i = 2, \dots, M, \\
 p_i^f(s+1) &= p_i + r_i Q\left(p_{i-1}^f(s+1), r_{i-1}^f(s+1), \right. \\
 &\quad \left. p_i^b(s+1), r_i^b(s+1), N_{i-1}\right), \\
 &\quad i = 2, \dots, M \quad (10)
 \end{aligned}$$

with function Q defined in (8) and the initial and boundary conditions given by

$$p_i^f(0) = p_i, \quad r_i^f(0) = r_i, \quad i = 2, \dots, M-1, \quad (11)$$

$$p_1^f(s) = p_1, \quad r_1^f(s) = r_1,$$

$$p_M^b(s) = p_M, \quad r_M^b(s) = r_M, \quad s = 1, 2, \dots \quad (12)$$

It has been shown in [11] that this recursive procedure is convergent and the following limits exist:

$$\begin{aligned}
 \lim_{s \rightarrow \infty} p_i^f(s) &=: p_i^f, & \lim_{s \rightarrow \infty} r_i^f(s) &=: r_i^f, \\
 \lim_{s \rightarrow \infty} p_i^b(s) &=: p_i^b, & \lim_{s \rightarrow \infty} r_i^b(s) &=: r_i^b, \\
 & & i &= 1, \dots, M. \quad (13)
 \end{aligned}$$

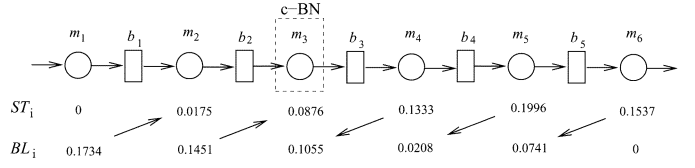


Fig. 2. Bottleneck identification rules.

Moreover

$$p_M^f = p_1^b, \quad r_M^f = r_1^b. \quad (14)$$

In terms of these limits, the production rate estimate of the M -machine line has been defined as follows:

$$\begin{aligned}
 PR &= e_M^f = e_1^b \\
 &= e_{i+1}^b \left[1 - Q\left(p_i^f, r_i^f, p_{i+1}^b, r_{i+1}^b, N_i\right) \right] \\
 &= e_i^f \left[1 - Q\left(p_{i+1}^b, r_{i+1}^b, p_i^f, r_i^f, N_i\right) \right] \\
 &\quad i = 2, \dots, M-1. \quad (15)
 \end{aligned}$$

Expressions (6)–(15) are used through out this paper for evaluating lean buffering of serial lines with exponential machines reliability.

B. Bottleneck Identification

In [12], the bottleneck, or more precisely, the c-bottleneck (c-BN) of a production line is defined as the machine i , such that

$$\frac{\partial PR}{\partial c_i} > \frac{\partial PR}{\partial c_j}, \quad \forall j \neq i. \quad (16)$$

where c_i is the number of parts the i th machine can produce per unit of time. Since in model (i)–(vi) all machines have the same c_i , (16) implies that the c-BN is a machine, which leads to the largest increase of the PR if its capacity, c_i , is increased (as compared with a similar increase of the capacity of any other machine in the system).

To outline a method of c-BN identification, we note that as it has been shown in [9] and [10], aggregation procedure (10)–(12) can be used to evaluate the probabilities of blockage BL_i , $i = 1, \dots, M-1$, and starvation ST_i , $i = 2, \dots, M$ of all machines in the system. Specifically, the estimates of these probabilities are

$$BL_i = e_i Q\left(p_{i+1}^b, r_{i+1}^b, p_i^f, r_i^f, N_i\right), \quad i = 1, \dots, M-1 \quad (17)$$

$$ST_i = e_i Q\left(p_{i-1}^f, r_{i-1}^f, p_i^b, r_i^b, N_{i-1}\right), \quad i = 2, \dots, M \quad (18)$$

where p_i^b , r_i^b , p_i^f , and r_i^f are defined by (12) and Q is defined by (8). Using these probabilities, the bottleneck machine is identified as follows [12].

Consider a production line and place the probabilities of starvations and blockages under each machine, as shown in Fig. 2.

Assign the arrows directed from one machine to another according to the following rule: If $BL_i > ST_{i+1}$, the arrow is pointing from m_i to m_{i+1} . If $BL_i < ST_{i+1}$, the arrow is pointing from m_{i+1} to m_i . Then, as it has been shown in [12], if there is a unique machine with no emanating arrows, it is the c-BN in the sense of (16). If there are multiple machines with no emanating arrows, the one with the largest severity is the primary bottleneck (PBN), where the severity is defined by

$$\begin{aligned} S_i &= (BL_{i-1} + ST_{i+1}) - (BL_i + ST_i), \\ & \quad i = 2, \dots, M-1 \\ S_1 &= ST_2 - BL_1 \\ S_M &= BL_{M-1} - ST_M. \end{aligned} \quad (19)$$

This method of BN identification is used in Section V for one of the recursive techniques for LLB evaluation.

C. LLB in Serial Lines With Identical Exponential Machines

A closed-form expression for LLB in two-machines lines with identical exponential machines has been derived in [1] to be

$$k_E(M=2) = \begin{cases} \frac{2e(E-e)}{1-E}, & \text{if } e < E \\ 0, & \text{otherwise} \end{cases}. \quad (20)$$

For $M > 2$, it has been shown in [1] that the following formula provides a sufficiently precise estimate for LLB:

$$\hat{k}_E(M > 2) = \begin{cases} \frac{e(1-\hat{Q})(e\hat{Q}+1-e)(e\hat{Q}+2-2e)(2-\hat{Q})}{\hat{Q}(2e-2e\hat{Q}+e\hat{Q}^2+\hat{Q}-2)} \ln \left(\frac{E-eE+eE\hat{Q}-1+e-2e\hat{Q}+e\hat{Q}^2+\hat{Q}}{(1-e-\hat{Q}+e\hat{Q})(E-1)} \right), & \text{if } e < E^{\frac{1}{M-1}} \\ 0, & \text{otherwise} \end{cases}, \quad (21)$$

where \hat{Q} is given by

$$\begin{aligned} \hat{Q} &= 1 - E^{\frac{1}{2}} \left[1 + \left(\frac{M-3}{M-1} \right)^{M/4} \right] \\ & \quad + \left(E^{\frac{1}{2}} \left[1 + \left(\frac{M-3}{M-1} \right)^{M/4} \right] - E^{\frac{M-2}{M-1}} \right) \\ & \quad \cdot \exp \left\{ - \left(\frac{E^{\frac{1}{M-1}} - e}{1-E} \right) \right\}. \end{aligned} \quad (22)$$

Expressions (20)–(22) are used in Section V for closed formula-based approaches to LLB evaluation in serial lines with nonidentical machines.

IV. TWO NONIDENTICAL MACHINE LINES

In the case of nonidentical two-machines lines, as it follows from (2) and (6), the equation that defines LBC, N_E , is

$$\begin{aligned} PR &= PR_\infty E = e_2 [1 - Q(p_1, r_1, p_2, r_2, N_E)] \\ &= e_2 \left[1 - \frac{(1-e_1)(1-\phi)}{1-\phi \exp -\beta N_E} \right] \end{aligned} \quad (23)$$

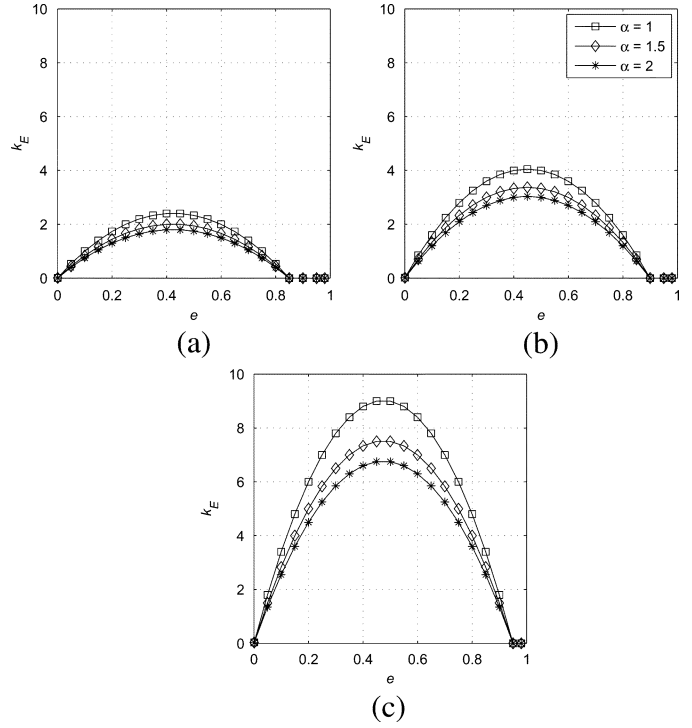


Fig. 3. Lean buffering as a function of identical machine efficiency. (a) $E = 0.85$. (b) $E = 0.90$. (c) $E = 0.95$.

where

$$PR_\infty = \min(e_1, e_2) \quad (24)$$

$$\phi = \frac{e_1(1-e_2)}{e_2(1-e_1)} \quad (25)$$

and

$$\beta = \frac{(r_1 + r_2 + p_1 + p_2)(p_1 r_2 - p_2 r_1)}{(r_1 + r_2)(p_1 + p_2)}. \quad (26)$$

From here we obtain the following.

Proposition 1: The LLB in serial lines defined by assumptions (i)–(vi) with $M = 2$ is given by

$$k_E(p_1, r_1, p_2, r_2) = \begin{cases} \frac{\min(r_1, r_2)}{\beta} \ln \left\{ \phi \frac{(e_2 - E PR_\infty)}{(e_1 - E PR_\infty)} \right\}, & \text{if } \max(e_1, e_2) < E \\ 0, & \text{otherwise} \end{cases} \quad (27)$$

where PR_∞ , ϕ , and β are defined by (24)–(26), respectively.

Proof: The first expression in the upper line of (27) follows immediately from (1)–(4). The condition $\max(e_1, e_2) < E$ ensures that this expression takes positive values for all positive p_i and r_i , $i = 1, 2$, and $0 < E < 1$.

Comparing (20) and (27), one can see that unlike the identical machine case, k_E in the nonidentical machine situation depends explicitly not only on the machines' efficiency but also on their uptimes and downtimes. To analyze this dependency, assume that $e_1 = e_2 = e$ and $r_1 = \alpha r_2$, where $\alpha \geq 1$. The resulting behavior of k_E as a function of e for various values of α and E is shown in Fig. 3. From this figure, we conclude the following.

- (α) For all α and E , the qualitative behavior of k_E as a function of e remains the same as in the identical machines case (see [1, Fig. 3(a)] and comments therein).

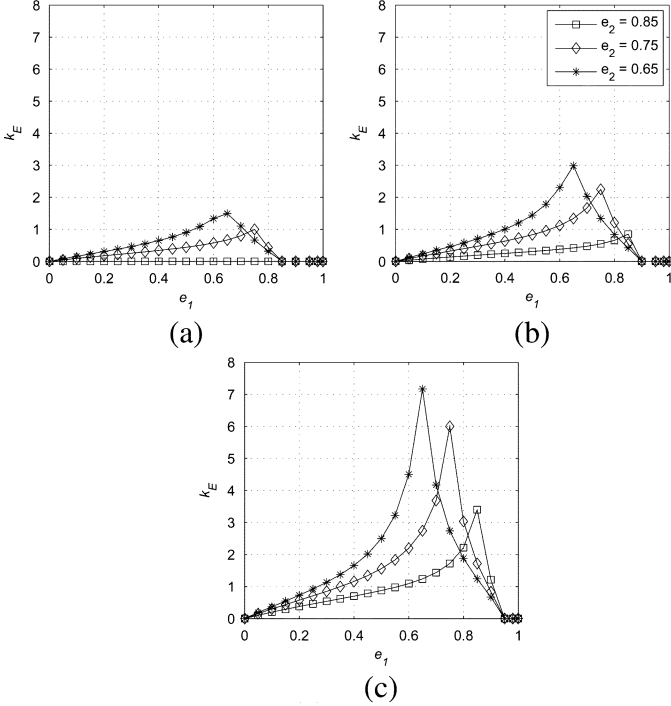


Fig. 4. Lean buffering as a function of the first machine efficiency. (a) $E = 0.85$. (b) $E = 0.90$. (c) $E = 0.95$.

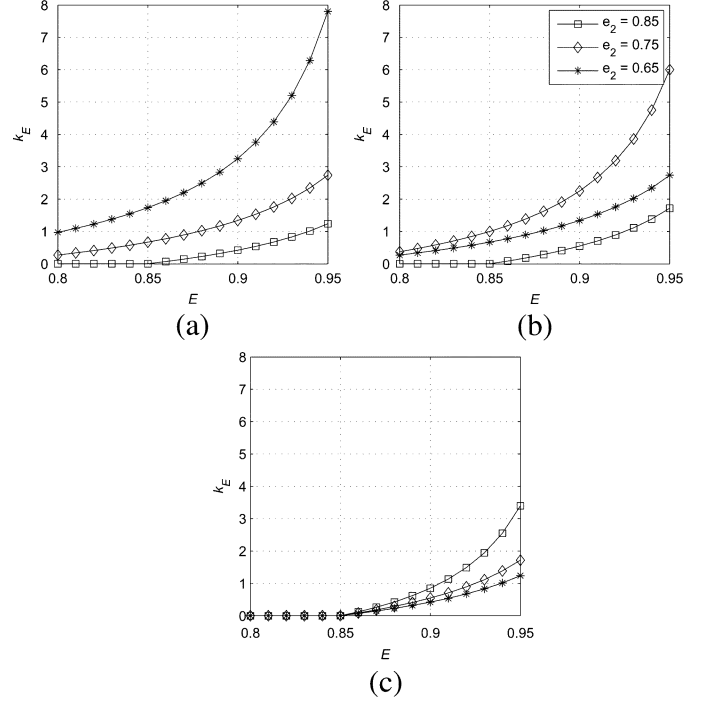


Fig. 5. Lean buffering as a function of line efficiency. (a) $e_1 = 0.65$. (b) $e_1 = 0.75$. (c) $e_1 = 0.85$.

- (β) However, larger α 's lead to smaller LLB and, consequently, smaller LBC. Since for each e , large α implies that one of the machines has shorter uptime and downtime, this leads to the conclusion that if the efficiency of both machines remains the same, it is beneficial to decrease the downtime of at least one machine, even at the expense of decreasing its uptime (e.g., for preventative maintenance).

Figs. 4 and 5 illustrates the behavior of k_E when the two machines have different efficiencies. Specifically, Fig. 4 characterizes k_E as a function of e_1 for various values of e_2 and E , while Fig. 5 shows k_E as a function of E for various e_1 and e_2 . In these figures, we let the average downtimes for both machines to be the same. The values for LLB are the same as long as downtimes of both machines are equal to each other and remain constant as other parameters change. From these figures, we conclude the following.

- (α) For e_2 sufficiently large, one downtime of LLB is acceptable for all values of e_1 and E .
- (β) For small e_2 , one downtime of LLB is acceptable only when e_1 is sufficiently large. For instance, if $e_2 = 0.75$, 1 downtime of LLB is sufficient only if $e_1 > 0.82$ for $E = 0.9$ and $e_1 > 0.90$ for $E > 0.95$.
- (γ) The maximum of k_E tends to take place when $e_1 = e_2$.

Intuitively, it is expected that the lean buffering in a line with machines $\{p_1, r_1, p_2, r_2\}$ is the same as in the reversed line, i.e., $\{p_2, r_2, p_1, r_1\}$. It turns out that this is indeed true as stated next.

Proposition 2: The lean buffering has the property of reversibility, i.e.,

$$k_E(p_1, r_1, p_2, r_2) = k_E(p_2, r_2, p_1, r_1). \quad (28)$$

Proof: Following immediately from (27) by observing that

$$\beta(p_1, r_1, p_2, r_2) = -\beta(p_2, r_2, p_1, r_1)$$

and

$$\phi(e_1, e_2) = 1/\phi(e_2, e_1).$$

In conclusion of this section, note that in terms of LBC, formula (27) can be rewritten as

$$N_E(p_1, r_1, p_2, r_2) = \begin{cases} \left\lceil \frac{1}{\beta} \ln \left\{ \phi \left(\frac{e_2 - EPR_\infty}{e_1 - EPR_\infty} \right) \right\} \right\rceil, & \text{if } \max(e_1, e_2) < E \\ 0, & \text{otherwise} \end{cases} \quad (29)$$

where $\lceil x \rceil$ denotes the smallest integer larger than x . This formula is used below for estimating LLB in serial lines with $M > 2$.

V. $M > 2$ -MACHINE LINES

Exact formulas for LLB or LBC in the case of $M > 2$ are all but impossible to derive [due to the complex structure of (10)]. Therefore, we limit our attention to estimates of $N_{i,E}$. These estimates are obtained using either closed formulas (29) and (20)–(22) or recursive calculations. Each of the approaches analyzed is described next.

A. Closed Formula Approaches

The following four methods have been investigated.

1. *Local pairwise approach.* Consider every pair of consecutive machines, m_i and m_{i+1} , $i = 1, \dots, M-1$, and select LBC using formula (29). This results in the sequence of buffer capacities denoted as

$$N_{1,E}^I, \dots, N_{M-1,E}^I.$$

Normalizing $N_{i,E}^I$ by $\max\{T_{\text{down},i}, i = 1, \dots, M\}$, results in

$$k_{1,E}^I, \dots, k_{M-1,E}^I.$$

II. *Global pairwise approach.* It is based on applying formula (29) to all possible pairs of machines (not necessarily consecutive) and then selecting the capacity of each buffer equal to the largest buffer obtained by this procedure. Clearly, this results in buffers of equal capacity, which is denoted as N_E^{II} . Normalizing this by $\max\{T_{\text{down},i}, i = 1, \dots, M\}$, results in k_E^{II} .

III. *Local upper bound approach.* Consider all pairs of consecutive machines, m_i and $m_{i+1}, i = 1, \dots, M - 1$, substitute each of them by a two-machine line with identical machines defined by

$$\hat{e}_i := \min\{e_i, e_{i+1}\}, i = 1, \dots, M - 1$$

$$\hat{r}_i := \min\{r_i, r_{i+1}\}, i = 1, \dots, M - 1$$

and

$$\hat{p}_i = \frac{\hat{r}_i}{\hat{e}_i} - \hat{r}_i, i = 1, \dots, M - 1$$

and select LBC using formula (20) with $e = \hat{e}_i, r = \hat{r}_i$, and $p = \hat{p}_i$. This results in the sequence of buffer capacities

$$N_{1,E}^{\text{III}}, \dots, N_{M-1,E}^{\text{III}}$$

Normalizing $N_{i,E}^{\text{III}}$ by $\max\{T_{\text{down},i}, i = 1, \dots, M\}$, results in

$$k_{1,E}^{\text{III}}, \dots, k_{M-1,E}^{\text{III}}$$

IV. *Global upper bound approach.* In stead of the original line, consider a line with all identical machines specified by

$$\hat{e} := \min\{e_1, e_2, \dots, e_M\}$$

$$\hat{r} := \min\{r_1, r_2, \dots, r_M\}$$

and

$$\hat{p} = \frac{\hat{r}}{\hat{e}} - \hat{r}$$

and select the buffer capacity, denoted as N_E^{IV} , using expressions (21) and (22). Due to the monotonicity of PR with respect to the machine efficiency and buffer capacity, this approach provides an upper bound of LBC

$$N_{i,E} \leq N_E^{\text{IV}}, i = 1, \dots, M - 1.$$

Normalizing N_E^{IV} by $\max\{T_{\text{down},i}, i = 1, \dots, M\}$, results in k_E^{IV} .

If the desired line efficiency for two-machine lines, involved in approaches I–III, were selected as E , the resulting efficiency of the M -machine line would be certainly less than E . To avoid this, the efficiency, E' , of each of the two-machine lines is calculated as follows: For a given M -machine line, find the buffer capacity using approach IV. Then, consider a two-machine line with identical machines, where each machine is defined by $\hat{e} = \min\{e_1, \dots, e_M\}, \hat{r} = \min\{r_1, \dots, r_M\}$, and $\hat{p} = (\hat{r}/\hat{e}) - \hat{r}$, and the buffer with the capacity as found above. Finally, calculate the production rate and the efficiency of this two-machine line and use it as E' in approaches I–III.

TABLE I
PERFORMANCE CHARACTERISTICS OF APPROACHES I–IV

Method	I	II	III	IV
k_{ave}^j	1.2 ± 0.01	5.0 ± 0.02	5.4 ± 0.03	9.5 ± 0.05
Δ^j	96.7	0.1	0.0	0.0

To analyze the performance of approaches I–IV, we considered 100 000 lines formed by selecting M, e_i , and $T_{\text{down},i}$ randomly and equiprobably from the sets

$$M \in \{4, 5, \dots, 30\} \tag{30}$$

$$0.70 \leq e \leq 0.97 \tag{31}$$

$$5 \leq T_{\text{down}} \leq 50. \tag{32}$$

The desired efficiency for each of these lines was also selected randomly and equiprobably from the set

$$0.80 \leq E \leq 0.98. \tag{33}$$

For each s th line thus formed, we calculated the vector of estimates of LLB

$$\mathbf{k}_s^j = \begin{bmatrix} k_{1,s}^j \\ k_{2,s}^j \\ \dots \\ k_{M-1,s}^j \end{bmatrix}, s = 1, \dots, 100\,000, j = \text{I, II, III, IV} \tag{34}$$

using the four approaches introduced above. The subscripts of $k_{i,s}^j$ represent i th buffer, $i \in \{1, \dots, M_s - 1\}$, of the s th line, $s \in \{1, 2, \dots, 100\,000\}$; the superscript $j \in \{\text{I, II, III, IV}\}$ represents the approach used for this calculation. In addition, we calculated the production rate, PR_s^j , and the efficiency, E_s^j , using expressions (10)–(15) and (2), respectively. The efficacy of approaches I–IV has been characterized by two metrics.

1) The average level of buffering per machine

$$k_{\text{ave}}^j = \frac{1}{S} \sum_{s=1}^S k_s^j \tag{35}$$

where $S = 100\,000$ and

$$k_s^j = \frac{1}{M_s - 1} \sum_{i=1}^{M_s - 1} k_{i,s}^j.$$

2) The frequency of E_s^j being less than the desired efficiency E_s

$$\Delta^j = \frac{1}{S} \sum_{s=1}^S Sg(E_s - E_s^j) \cdot 100\% \tag{36}$$

where $S = 100\,000$ and

$$Sg(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}.$$

The results are given in Table I along with the 95% confidence intervals for LLB. (Note that the confidence intervals for Δ 's are not included since they would be meaningless, given that the Δ 's of Table I and of all subsequent tables are either close to 100% or 0%.) Clearly, Approach I leads to the smallest

TABLE II
EFFECT OF M ON THE PERFORMANCE OF APPROACHES I–IV.
(a) $M \in \{5, 10, 15, 20\}$. (b) $M \in \{25, 30, 50, 80\}$

(a)				
M	5	10	15	20
k_{ave}^I	1.2 ± 0.02	1.2 ± 0.02	1.2 ± 0.02	1.2 ± 0.02
k_{ave}^{II}	3.2 ± 0.05	4.5 ± 0.07	5.1 ± 0.07	5.5 ± 0.08
k_{ave}^{III}	5.3 ± 0.09	5.5 ± 0.09	5.5 ± 0.09	5.4 ± 0.09
k_{ave}^{IV}	7.8 ± 0.12	9.3 ± 0.14	9.8 ± 0.15	10.0 ± 0.15
Δ^I	88.3	96.2	98.2	99.0
Δ^{II}	0.6	0.0	0.0	0.0
Δ^{III}	0.0	0.0	0.0	0.0
Δ^{IV}	0.0	0.0	0.0	0.0

(b)				
M	25	30	50	80
k_{ave}^I	1.2 ± 0.02	1.2 ± 0.02	1.2 ± 0.02	1.2 ± 0.02
k_{ave}^{II}	5.7 ± 0.08	5.9 ± 0.08	6.4 ± 0.09	6.6 ± 0.09
k_{ave}^{III}	5.4 ± 0.08	5.5 ± 0.09	5.4 ± 0.08	5.4 ± 0.08
k_{ave}^{IV}	10.0 ± 0.15	10.3 ± 0.15	10.4 ± 0.15	10.4 ± 0.15
Δ^I	99.4	99.5	99.9	99.9
Δ^{II}	0.0	0.0	0.0	0.0
Δ^{III}	0.0	0.0	0.0	0.0
Δ^{IV}	0.0	0.0	0.0	0.0

average level of buffering but, unfortunately, almost always results in line efficiency less than desired. Thus, a “local” selection of LLB (i.e., based on the two machines surrounding the buffer) is unacceptable. Approaches II and III provide line efficiency less than desired in only a small fraction of cases and result in the average buffer capacity 4–5 times larger than Approach I. Approach IV, as expected, always guarantees the desired performance but requires the largest buffering.

To further differentiate between the four approaches, we considered their performance as a function of M . To accomplish this, we formed 10 000 lines for each $M \in \{5, 10, 15, 20, 25, 30, 50, 80\}$ by selecting e_i 's, $T_{down,i}$'s, and E 's randomly and equiprobably from sets (31)–(33), respectively. For each of these lines, we calculated buffer levels using approaches I–IV and evaluated the performance metrics (35) and (36). The results are shown in Table II along with 95% confidence intervals. Examining these data, we conclude the following.

- (α) The local pairwise approach in almost all cases leads to a lower line efficiency than desired.
- (β) The global pairwise approach results in good performance from the point of view of both k_{ave} and Δ . For $M \leq 15$, it outperforms Approach III from the point of view of k_{ave} . However, k_{ave} increases as M increases.
- (γ) The local upper bound approach is less sensitive to M and outperforms Approach II for $M > 15$.
- (δ) The global upper bound approach substantially overestimates LLB.

Based on the above, it is recommended to use the global pairwise approach in systems with $M \leq 15$ and local upper bound approach in systems with $M > 15$.

B. Recursive Approaches

The following two recursive methods have been investigated.

- V. *Full search approach.* Start from all buffers of capacity 1. Increase the capacity of the first buffer by 1 and, using the aggregation procedure (10)–(15), calculate the pro-

duction rate of the system. Return the first buffer capacity to its initial value, increase the second buffer capacity by 1 and calculate the resulting production rate. Repeat the same procedure for all buffers, determine the buffer that leads to the largest production rate, and permanently increase its capacity by 1. Repeat the same procedure until the desired line efficiency is reached. This will result in the sequence of buffer capacities

$$N_{1,E}^V, \dots, N_{M-1,E}^V.$$

Normalizing $N_{i,E}^V$ by $\max\{T_{down,i}, i = 1, \dots, M\}$, results in

$$k_{1,E}^V, \dots, k_{M-1,E}^V.$$

- VI. *Bottleneck-based approach.* Consider a production line with the buffer capacity calculated according to Approach I but rounded down in formula (29) rather than up. Although being relatively small, this buffering often leads to line efficiency less than desired. Therefore, to improve the line efficiency, increase the buffering according to the following procedure: Using the technique described in Section III-B, identify the bottleneck (or, when applicable, primary bottleneck) and increase the capacity of both buffers surrounding this machine by 1. Repeat this procedure until the desired line efficiency is reached. This will result in the sequence of buffer capacities denoted as

$$N_{1,E}^{VI}, \dots, N_{M-1,E}^{VI}.$$

Normalizing $N_{i,E}^{VI}$ by $\max\{T_{down,i}, i = 1, \dots, M\}$, results in

$$k_{1,E}^{VI}, \dots, k_{M-1,E}^{VI}.$$

Clearly, Approach V gives in a smaller buffer capacity than Approach VI. However, the latter might require a shorter computation time than the former. Therefore, in order to compare approaches V and VI, the computation time should also be taken into account. This additional performance metric is defined as the total computer time necessary to carry out the computation:

$$t^j = t_{end}^j - t_{start}^j \quad (37)$$

where t_{start}^j and t_{end}^j are the times (in milliseconds) of the beginning and the end of the computation.

Based on performance metrics (35)–(37), we compared approaches I–VI using the production systems generated by selecting e_i 's and $T_{down,i}$'s randomly and equiprobably from sets (31) and (32). The results are shown in Tables III–V along with 95% confidence intervals. Specifically, Tables III and IV present the results obtained using 5000 randomly generated 5- and 10-machine lines, respectively, while Table V is based on the analysis of 2000 randomly generated 15-machine lines. Examining these data, we conclude the following.

- (α) The full search approach, as expected, results in the smallest level of buffering and the longest calculation time.
- (β) Approaches II–IV, being based on closed-form expressions, are practically instantaneous but lead to the average buffering 2–10 times larger than that of Approach V.

TABLE III
PERFORMANCE CHARACTERISTICS OF APPROACHES I–VI IN 5-MACHINE LINES.
(a) $E = 0.80$. (b) $E = 0.85$. (c) $E = 0.90$. (d) $E = 0.95$

(a)			
Approaches	I	II	III
k_{ave}^j	0.4 ± 0.01	1.1 ± 0.01	1.6 ± 0.01
Δ^j	92.4	2.3	0.0
t^j	0	0	0
Approaches	IV	V	VI
k_{ave}^j	2.7 ± 0.01	0.5 ± 0.01	0.6 ± 0.01
Δ^j	0.0	0.0	0.0
t^j	0	132	8

(b)			
Approaches	I	II	III
k_{ave}^j	0.6 ± 0.01	1.7 ± 0.02	2.4 ± 0.02
Δ^j	93.0	0.8	0.0
t^j	0	0	0
Approaches	IV	V	VI
k_{ave}^j	3.8 ± 0.01	0.7 ± 0.01	0.9 ± 0.01
Δ^j	0.0	0.0	0.0
t^j	0	211	12

(c)			
Approaches	I	II	III
k_{ave}^j	1.0 ± 0.02	2.6 ± 0.03	4.0 ± 0.03
Δ^j	90.2	0.2	0.0
t^j	0	0	0
Approaches	IV	V	VI
k_{ave}^j	6.0 ± 0.02	1.1 ± 0.01	1.4 ± 0.02
Δ^j	0.0	0.0	0.0
t^j	0	355	17

(d)			
Approaches	I	II	III
k_{ave}^j	2.0 ± 0.03	5.0 ± 0.06	8.6 ± 0.05
Δ^j	81.2	0.0	0.0
t^j	0	0	0
Approaches	IV	V	VI
k_{ave}^j	12.6 ± 0.04	1.9 ± 0.03	2.4 ± 0.03
Δ^j	0.0	0.0	0.0
t^j	0	679	22

- (γ) Approach VI provides a good tradeoff between the calculation time and level of buffering. It is up to two orders of magnitude faster than Approach V and results in the average level of buffering slightly larger than that of Approach V (about 20% difference). It is, of course, slower than approaches I–IV but gives buffering 2–6 times smaller than approaches II–IV.

C. Illustrative Examples

To illustrate particular cases of lean buffering designed using approaches I–VI, we provide several examples.

Consider the four production lines with machines specified in Table VI along with the desired line efficiency. The estimates of LLB for each of these lines, calculated using approaches I–VI, are shown in Table VII–X; these tables include also the resulting line efficiency, E^j . These examples clearly show the efficacy of the bottleneck approach, which is based

TABLE IV
PERFORMANCE CHARACTERISTICS OF APPROACHES I–VI IN 10-MACHINE LINES. (a) $E \in [0.80, 0.89]$. (b) $E \in [0.89, 0.98]$

(a)			
Approaches	I	II	III
k_{ave}^j	0.6 ± 0.01	2.4 ± 0.02	2.6 ± 0.02
Δ^j	99.7	0.0	0.0
t^j	1	1	1
Approaches	IV	V	VI
k_{ave}^j	4.7 ± 0.02	0.8 ± 0.01	1.0 ± 0.01
Δ^j	0.0	0.0	0.0
t^j	1	3,983	103

(b)			
Approaches	I	II	III
k_{ave}^j	1.9 ± 0.03	6.7 ± 0.10	8.4 ± 0.14
Δ^j	93.3	0.0	0.0
t^j	1	1	1
Approaches	IV	V	VI
k_{ave}^j	14.1 ± 0.21	1.7 ± 0.02	2.3 ± 0.03
Δ^j	0.0	0.0	0.0
t^j	1	11,576	147

TABLE V
PERFORMANCE CHARACTERISTICS OF APPROACHES I–VI IN 15-MACHINE LINES. (a) $E \in [0.80, 0.89]$. (b) $E \in [0.89, 0.98]$

(a)			
Approaches	I	II	III
k_{ave}^j	0.6 ± 0.01	2.8 ± 0.03	2.6 ± 0.03
Δ^j	100.0	0.0	0.0
t^j	2	2	2
Approaches	IV	V	VI
k_{ave}^j	5.0 ± 0.04	0.9 ± 0.01	1.1 ± 0.01
Δ^j	0.0	0.0	0.0
t^j	2	24,403	375

(b)			
Approaches	I	II	III
k_{ave}^j	1.9 ± 0.03	6.7 ± 0.02	8.4 ± 0.03
Δ^j	92.8	0.0	0.0
t^j	2	2	2
Approaches	IV	V	VI
k_{ave}^j	14.1 ± 0.02	1.7 ± 0.02	2.3 ± 0.02
Δ^j	0.0	0.0	0.0
t^j	2	71,300	514

TABLE VI
DESIRED LINE EFFICIENCIES AND MACHINE PARAMETERS OF CASE STUDIES

Line	E^j	e_1	e_2	e_3	e_4	e_5
1	0.80	0.83	0.88	0.71	0.74	0.90
2	0.85	0.97	0.76	0.79	0.75	0.90
3	0.90	0.91	0.87	0.76	0.84	0.78
4	0.95	0.79	0.88	0.96	0.95	0.81

Line	$T_{down,1}$	$T_{down,2}$	$T_{down,3}$	$T_{down,4}$	$T_{down,5}$
1	22	39	17	23	28
2	22	24	49	47	30
3	33	20	31	27	29
4	32	15	35	37	19

on the closed-formula (27) and the BN identification method described in Section III.

TABLE VII
LLB ESTIMATES FOR LINE 1

Buffer	b_1^j	b_2^j	b_3^j	b_4^j	E^j
<i>Desired</i>					0.80
k_i^I	0.3	0.2	1.1	0.1	0.71
k_i^{II}	1.9	1.9	1.1	1.4	0.87
k_i^{III}	1.2	2.9	1.7	1.8	0.90
k_i^{IV}	2.9	2.9	2.9	2.9	0.95
k_i^V	0.5	1.0	1.4	0.3	0.80
k_i^{VI}	0.3	1.1	1.9	0.1	0.80

TABLE VIII
LLB ESTIMATES FOR LINE 2

Buffer	b_1^j	b_2^j	b_3^j	b_4^j	E^j
<i>Desired</i>					0.85
k_i^I	0.0	1.7	2.1	0.2	0.81
k_i^{II}	1.2	2.6	2.6	2.4	0.89
k_i^{III}	1.8	3.7	3.8	3.6	0.93
k_i^{IV}	3.8	3.8	3.8	3.8	0.93
k_i^V	0.2	2.1	2.3	0.8	0.85
k_i^{VI}	0.0	2.0	2.8	0.7	0.85

TABLE IX
LLB ESTIMATES FOR LINE 3

Buffer	b_1^j	b_2^j	b_3^j	b_4^j	E^j
<i>Desired</i>					0.90
k_i^I	0.8	1.0	1.8	1.9	0.84
k_i^{II}	4.4	4.1	4.1	3.8	0.95
k_i^{III}	2.7	5.4	5.4	4.8	0.97
k_i^{IV}	5.9	5.9	5.9	5.9	0.97
k_i^V	0.7	2.2	3.0	2.5	0.90
k_i^{VI}	0.8	2.4	3.3	1.9	0.90

TABLE X
LLB ESTIMATES FOR LINE 4

Buffer	b_1^j	b_2^j	b_3^j	b_4^j	E^j
<i>Desired</i>					0.95
k_i^I	1.6	0.4	0.7	0.6	0.82
k_i^{II}	5.5	6.1	6.4	6.4	1.00
k_i^{III}	9.4	6.1	1.5	10.3	1.00
k_i^{IV}	10.9	10.9	10.9	10.9	1.00
k_i^V	3.5	2.8	1.5	3.2	0.95
k_i^{VI}	2.9	1.8	0.8	6.4	0.95

VI. CONCLUSION

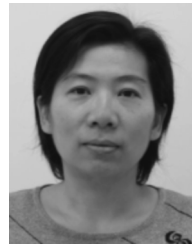
This paper analyzes the lean level of buffering for non-identical exponential machines in a serial production line. Closed-form expressions have been derived for two-machine lines. For longer lines, several approximation methods have been explored. Some of them utilize closed formulas mentioned above, while others are based on recursions. As a result, it has been shown that

- if closed formulas are to be used for selecting lean buffering, the global pairwise and the local upper bound approaches are recommended for $M \leq 15$ and $M > 15$, respectively;
- if recursive approaches are to be employed, the bottleneck-based approach is preferred.

Future research on the topic of lean buffering will include assembly systems and asynchronous production lines.

REFERENCES

- [1] E. Enginarlar, J. Li, and S. M. Meerkov, "How lean can lean buffers be?," *IIE Transactions*, vol. 37, pp. 333–342, 2005.
- [2] S. B. Gershwin and Y. Goldis, "Efficient algorithms for transfer line design," Lab. Manufacturing and Productivity, MIT, Cambridge, MA, Rep. LMP-95-005, 1995.
- [3] H. Yamashita and T. Altiok, "Buffer capacity allocation for a desired throughput of production lines," *IIE Transactions*, vol. 30, pp. 883–891, 1998.
- [4] S. B. Gershwin and J. E. Schor, "Efficient algorithms for buffer space allocation," *Ann. Oper. Res.*, vol. 93, pp. 117–144, 2000.
- [5] J. MacGregor Smith and F. R. B. Cruz, "The buffer allocation problem for general finite buffer queueing networks," *IIE Transactions*, vol. 37, pp. 343–365, 2005.
- [6] J. A. Buzacott, "Automatic transfer lines with buffer stocks," *Int. J. Prod. Res.*, vol. 5, pp. 183–200, 1967.
- [7] J. O. McClain, R. Conway, W. Maxwell, and L. J. Thomas, "The role of work-in-process inventory in serial production lines," *Oper. Res.*, vol. 36, pp. 229–241, 1988.
- [8] E. Enginarlar, J. Li, S. M. Meerkov, and R. Q. Zhang, "Buffer capacity for accommodating machine downtime in serial production lines," *Int. J. Prod. Res.*, vol. 40, pp. 601–624, 2002.
- [9] S.-Y. Chiang, C.-T. Kuo, and S. M. Meerkov, "Bottlenecks in Markovian production lines: A systems approach," *IEEE Trans. Robot. Autom.*, vol. 14, pp. 352–359, 1998.
- [10] S.-Y. Chiang, C.-T. Kuo, and S. M. Meerkov, "DT-bottlenecks in serial production lines: Theory and application," *IEEE Trans. Robot. Autom.*, vol. 16, pp. 567–580, 2000.
- [11] J. Li, "Performance analysis of production systems with rework loops," *IIE Transactions*, vol. 36, pp. 755–765, 2004.
- [12] S.-Y. Chiang, C.-T. Kuo, and S. M. Meerkov, "c-bottlenecks in serial production lines: Identification and application," *Mathematical Problems in Engineering*, vol. 7, pp. 543–578, 2001.



Shu-Yin Chiang received the B.S. degree from the Tatung Institute of Technology, Taipei, Taiwan, R.O.C., in 1990, and the M.S. and Ph.D. degrees from the University of Michigan, Ann Arbor, in 1994 and 1999, respectively, all in electrical engineering.

She is currently a Chair and Assistant Professor in the Department of Information and Telecommunications Engineering, Ming-Chuan University, Taiwan, R.O.C. Her research interests include manufacturing, communication networks, and wireless communication.



Alexander Hu received the B.S. degree in electrical engineering from Northwestern University, Evanston, IL, in 2004 and the M.S. degree in electrical engineering systems from the University of Michigan, Ann Arbor, in 2005.

He is currently an analyst for Capital One. His interests include analyzing manufacturing systems, credit risk, and equities investments.



Semyon M. Meerkov (M'78–SM'83–F'90) received the M.S.E.E. degree from the Polytechnic of Kharkov, Kharkov, Ukraine, in 1962 and the Ph.D. degree in systems science from the Institute of Control Sciences, Moscow, Russia, in 1966.

He was with the Institute of Control Sciences until 1977. From 1979 to 1984, he was with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL. Since 1984, he has been a Professor at the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. He has held visiting positions at the University of California, Los Angeles (UCLA) (1978–1979), Stanford University (1991), and the Lady Davis Visiting Professorship at the Technion, Israel (1997–1998). He is Editor-in-Chief of *Mathematical Problems in Engineering*, Department Editor for Manufacturing Systems of the *IIE Transactions*, and associate editor of several other journals. His research interests are in systems and control with applications to production systems and communication networks.