IIE Transactions (2005) **37**, 333–342
Copyright © "IIE"
ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/07408170590916878

How lean can lean buffers be?

EMRE ENGINARLAR[1], JINGSHAN LI[2] and SEMYON M. MEERKOV[3,*]

[1] *Decision Applications Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[2] *Manufacturing Systems Research Laboratory, GM Research and Development Center, Warren, MI 48090-9055, USA*
[3] *Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122, USA*
E-mail: smm@eecs.umich.edu

Received April 2003 and accepted October 2003

This paper provides a quantitative characterization of the smallest, i.e., lean, buffer capacity necessary and sufficient to attain a desired throughput in serial production lines with identical exponential machines. The development is carried out in terms of normalized buffer capacity and production line efficiency. The smallest normalized buffer capacity required to ensure the desired line efficiency is referred to as the Lean Level of Buffering (LLB). Exact formulas for the LLB in two- and three-machine lines are presented and an approximate expression for the LLB in lines with more than three machines is derived. Along with these analytical results, several qualitative insights into the nature of lean buffering in serial production lines are presented.

1. Introduction

Production lines with unreliable machines usually contain in-process buffers intended to decouple the machines and thereby reduce their mutual interference due to breakdowns. In the modern manufacturing environment, the buffers should be as small as possible, i.e., lean. But how lean can lean buffers be? In other words, how small can in-process buffers be to ensure the desired production rate of the overall system? The purpose of this paper is to provide a quantitative answer to this question.

To accomplish this, we consider serial production lines with unreliable machines and exponentially distributed up- and downtime (see Section 3 for precise model and problem formulation). The capacity of the buffer is quantified in units of average downtime. For instance, "k-downtime buffer" denotes the capacity of a buffer capable of storing the number of parts produced during k average downtimes. The number k is referred to as the *level of buffering*, LB. The overall *line efficiency*, E, is quantified as the ratio of the production rate of the line with LB equal to k to that of the line with an infinite LB. The question addressed is: how small can k be so that the desired line efficiency E is achieved? The smallest, i.e., lean, k that guarantees E is denoted as k_E and is referred to as the *lean level of buffering*, LLB.

In this paper, we develop quantitative methods for calculating lean level of buffering in serial production lines with machines having exponentially distributed up- and downtime. More specifically, we derive closed formulas, which quantify LLB as a function of system parameters.

Along with quantitative results, the paper provides several insights concerning lean buffering in serial lines with identical exponential machines. These include:

- LLB depends on three variables: (i) line efficiency, E; (ii) the number of machines in the system, M; and (iii) machine efficiency, e (defined as the ratio of the average uptime to the sum of the average up- and downtime).
- Lean systems need no buffering if and only if $e \geq E^{\frac{1}{M-1}}$.
- LLB is a concave function of the machine efficiency, e, reaching its maximum at around $e = 0.5E^{\frac{1}{M-1}}$.
- LLB is a monotonically increasing function of the line efficiency, E, with increasing growth rate, approaching infinity as $E \to 1$.
- LLB is a monotonically increasing function of the number of machines in the system, M, but with a decreasing growth rate. As a result, LLB for 15-machine lines is practically sufficient for lines with any $M > 15$.

The outline of this paper is as follows: Section 2 describes the related literature. In Section 3, the model of production systems under consideration is introduced and the problem of selecting LLB is formulated. The method of analysis is described in Section 4. Sections 5–7 present the main results. Discussions and extensions are in Section 8. The conclusions are formulated in Section 9. The proofs are given in the Appendix.

2. Related literature

Buffer capacity allocation in production lines has been studied quantitatively for over 50 years and numerous publications are available. The remarks below are intended to place the current research in the framework of this literature.

*Corresponding author

0740-817X © 2005 "IIE"

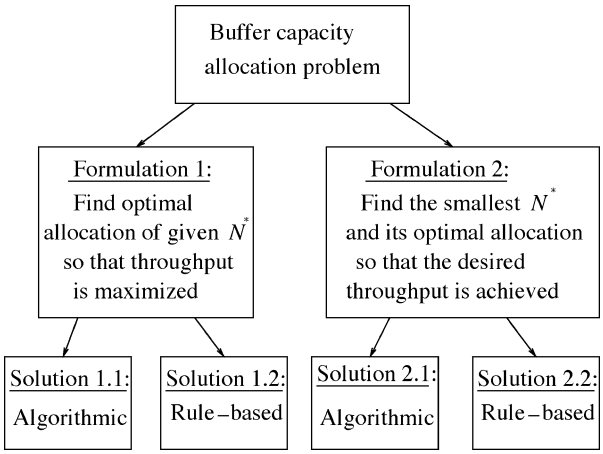

Fig. 1. Classification of the literature on the buffer capacity allocation problem.

The literature on in-process buffer capacity allocation can be classified as shown in Fig. 1. There are two different formulations. Formulation 1 seeks the optimal allocation of an *a-priori* given total buffer capacity, N^*, so that throughput is maximized. Formulation 2 seeks the smallest N^* and its allocation so that an *a-priori* given desired throughput is achieved (if at all possible). Each of these formulations may have two methods of solution: (i) algorithmic; and (ii) rule-based. Algorithmic methods lead to a computer code that provides a solution to a corresponding formulation. Rule-based methods give simple rules for either the best or good (i.e., near-optimal) solution of each formulation. Representative publications on the algorithmic solution of formulation 1 include Soyster *et al.* (1979), Altiok and Stidham (1983), Caramanis (1987), Chow (1987), Yamashita and Suzuki (1987), Smith and Daskalaki (1988), Jafari and Shantikumar (1989), Park (1993), Plambeck *et al.* (1993), Liu and Lin (1994), Seong *et al.* (1995), Harris and Powell (1999), and Tempelmeier (2003); the rule-based solutions are discussed in Freeman (1964), Sheskin (1976), Goldratt and Cox (1986), Hillier and So (1991a), Powell (1994), Jacobs and Meerkov (1995), Powell and Pyke (1996), and Chiang *et al.* (2000), respectively.

Formulation 2 has received much less attention. In fact, only a few recent papers propose its algorithmic solution (Gershwin and Goldis, 1995; Yamashita and Altiok, 1998; Gershwin and Schor, 2000). Rule-based approaches have been discussed in Buzacott (1967), McClain *et al.* (1988), and Enginarlar *et al.* (2002). Buzacott (1967) associates buffer capacity allocation with the average downtime and states that buffering beyond five-downtime can hardly be justified. McClain *et al.* (1988) also connect buffer allocation with downtime. They show that one-downtime buffering is sufficient to regain about 50% of production losses if the downtime is deterministic. They suggest that random (exponentially distributed) downtime may require twice this capacity to result in comparable gains. Enginarlar *et al.* (2002) provide an analytical expression for *LLB* in two-machine lines and use a numerical procedure to evaluate *LLB* in longer lines. The current paper is intended to provide additional results in the direction established by Buzacott (1967), McClain *et al.* (1988), and Enginarlar *et al.* (2002) by deriving closed formulas for *LLB* in serial lines with any finite number of exponential machines and providing several insights into the nature of lean buffering in manufacturing systems.

Another line of research on buffer capacity allocation is related to the so-called storage bowl phenomenon (Hillier and So, 1991b). According to this phenomenon, more buffering should be assigned to middle machines in production lines with identical machines. It can be shown, however, that unbalancing the buffering results in only 1–3% of throughput improvement, if at all. Since this improvement is quite small, the present study does not consider bowl-type storage allocations and assigns equal capacity to all buffers if the machines are identical.

3. Problem formulation

3.1. *Model*

The block diagram of the production system considered in this paper is shown in Fig. 2, where the circles represent the machines and the rectangles are the buffers. Although the development reported here can be carried out using various assumptions on the machines and buffers, and similar results can be obtained, we adopt the so-called flow (Gershwin, 1994) model with deterministic processing time, where the material processed is viewed as the flow of a fluid and processing of one part requires a fixed time (referred to as the machine *cycle time*). This model is appropriate when the average up- and downtime of the machines are much larger than the cycle times. This often takes place in machining and assembly lines in large volume manufacturing systems.

Specifically, we define the production line under consideration by the following assumptions:

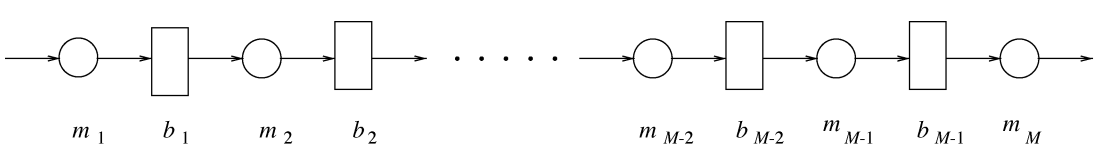

Fig. 2. Serial production line.

1. Each machine $m_i,\ i = 1, \ldots, M$, has two states: up and down. When up, the machine is capable of processing one part per cycle time. The cycle times of all machines are the same.
2. The up- and downtimes of each machine are random variables distributed exponentially with breakdown and repair rates p and r, respectively and independently of the buffer occupancy. Both p and r are in units of 1/(cycle time).
3. Each buffer is capable of storing N parts, $0 \le N \le \infty$. The capacity $N = 0$ implies that the amount of work produced during infinitesimal interval δt (in units of cycle time) is immediately transferred to the subsequent machine, if it is up. The capacity $N > 0$ implies that the work processed by a machine during δt is placed in the subsequent buffer, if it is not full.
4. Machine $m_i,\ i = 2, \ldots, M$, is starved at time t if it is up at time t, buffer b_{i-1} is empty at time t and m_{i-1} does not place any work in this buffer at time t. Machine m_1 cannot be starved.
5. Machine $m_i,\ i = 1, \ldots, M-1$, is blocked at time t if it is up at time t, buffer b_i is full at time t and m_{i+1} fails to take any work from this buffer at time t. Machine m_M cannot be blocked.

Remarks:

- Assumptions (1)–(3) imply that all machines are identical and all buffers are of equal capacity. We make these assumptions in order to provide a compact characterization of lean buffering.
- Assumption (2) implies, in particular, that the time-dependent failures, rather than the operation-dependent failures, are considered. Although the latter are, perhaps, more prevalent in practice, the former lead to a simpler analysis. Since for most practical values of machine and buffer parameters the two failure modes exhibit similar performance (Buzacott and Shantikumar, 1993), we adopt here the simpler version.
- Assumption (2) implies, in particular, that the average up- and downtime of the machines (T_{up} and T_{down}) are measured in units of cycle time, and the efficiency of each machine in isolation is:

$$e = \frac{r}{r+p} = \frac{T_{\text{up}}}{T_{\text{up}} + T_{\text{down}}}. \tag{1}$$

3.2. *Parametrization and problem formulation*

As was mentioned in the Introduction, it is convenient to represent the buffer capacity in units of average downtime. This representation is referred to as the level of buffering and is defined by:

$$k = \frac{N}{T_{\text{down}}}. \tag{2}$$

Another normalization used in this work is referred to as the line efficiency, E. It is introduced as follows. Let PR_k denote the production rate (i.e., the average number of parts produced by the last machine per cycle time) of the line when the level of buffering is $0 \le k < \infty$. Let PR_∞ denote the production rate of the same line with infinite capacity buffering. Then:

$$E = \frac{PR_k}{PR_\infty}. \tag{3}$$

Obviously, $0 < E < 1$. The smallest LB necessary to ensure the desired E is referred to as the lean level of buffering and denoted as k_E.

The problem addressed in this work is: *given the production line defined by assumptions 1–5, develop a method for calculating k_E.*

A solution to this problem is given in Sections 5–7.

4. Approach

The analysis described in this work is based on a technique for production rate evaluation developed in Chiang *et al.* (2000). This technique is applicable to serial lines defined by assumptions (1–5) even if the machines and buffers are non-identical, i.e., each machine m_i has its up- and downtime distributed exponentially with parameters p_i and r_i, respectively, $i = 1, \ldots, M$, and each buffer b_i has the capacity $N_i,\ i = 1, \ldots, M-1$. This technique is briefly outlined below (see Chiang *et al.* (2000) for details).

In the case of two-machine lines, the production rate, PR, is calculated according to the following formula:

$$\begin{aligned} PR &= e_2[1 - Q(p_1, r_1, p_2, r_2, N_1)], \\ &= e_1[1 - Q(p_2, r_2, p_1, r_1, N_1)], \end{aligned} \tag{4}$$

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

$$e_i = \frac{r_i}{r_i + p_i}, \tag{5}$$

and function $Q(p_1, r_1, p_2, r_2, N_1)$ is given by

$$Q(p_1, r_1, p_2, r_2, N_1) = \begin{cases} \dfrac{(1-e_1)(1-\phi)}{1 - \phi \exp\{-\beta N_1\}}, & \text{if } \dfrac{p_1}{r_1} \ne \dfrac{p_2}{r_2}, \\[2ex] \dfrac{p_1(p_1+p_2)(r_1+r_2)}{(p_1+r_1)[(p_1+p_2)(r_1+r_2) + p_2 r_1 (p_1+p_2+r_1+r_2)N_1]}, & \text{if } \dfrac{p_1}{r_1} = \dfrac{p_2}{r_2}, \end{cases} \tag{6}$$

$$\begin{aligned} \phi &= \frac{e_1(1-e_2)}{e_2(1-e_1)}, \\ \beta &= \frac{(r_1+r_2+p_1+p_2)(p_1 r_2 - p_2 r_1)}{(r_1+r_2)(p_1+p_2)}. \end{aligned} \tag{7}$$

Function $Q(p_2, r_2, p_1, r_1, N_1)$ is obtained from $Q(p_1, r_1, p_2, r_2, N_1)$ by replacing all quantities with subscript "1" (respectively, "2") by the similar quantities with subscript "2" (respectively, "1"). As is indicated in Chiang *et al.* (2000), function $Q(p_1, r_1, p_2, r_2, N_1)$ represents the probability that the buffer is empty whereas function $Q(p_2, r_2, p_1, r_1, N_1)$ gives the probability that the buffer is full.

For $M >$ two-machine lines, PR is evaluated using an iterative aggregation procedure based on Equations (4)–(7). This procedure consists of the so-called backward and forward aggregation. In the backward aggregation, the last two machines are aggregated into a single machine m^{b}_{M-1} defined by parameters p^{b}_{M-1} and r^{b}_{M-1}. Then m^{b}_{M-1} is aggregated with m_{M-2} to result in m^{b}_{M-2}, which is then aggregated with m_{M-3} to give m^{b}_{M-3}, and so on until all the machines are aggregated into m^{b}_1. In the forward aggregation, m_1 is aggregated with m^{b}_2 to produce m^{f}_2, which is then aggregated with m^{b}_3 to give m^{f}_3 and so on until all the machines are aggregated into m^{f}_M. Then the process is repeated anew. Formally, this recursive procedure can be represented as follows:

$$
\begin{aligned}
p^{\mathrm{b}}_i(s+1) &= \frac{p}{1 - Q\big(p^{\mathrm{b}}_{i+1}(s+1), r^{\mathrm{b}}_{i+1}(s+1), p^{\mathrm{f}}_i(s), r^{\mathrm{f}}_i(s), N_i\big)}, \quad 1 \le i \le M-1,\\
r^{\mathrm{b}}_i(s+1) &= 1 \bigg/ \left(\frac{Q\big(p^{\mathrm{b}}_{i+1}(s+1), r^{\mathrm{b}}_{i+1}(s+1), p^{\mathrm{f}}_i(s), r^{\mathrm{f}}_i(s), N_i\big)}{p} + \frac{1}{r} \right), \quad 1 \le i \le M-1,\\
p^{\mathrm{f}}_i(s+1) &= \frac{p}{1 - Q\big(p^{\mathrm{b}}_{i-1}(s+1), r^{\mathrm{f}}_{i-1}(s+1), p^{\mathrm{b}}_i(s+1), r^{\mathrm{b}}_i(s+1), N_{i-1}\big)}, \quad 2 \le i \le M,\\
r^{\mathrm{f}}_i(s+1) &= 1 \bigg/ \left(\frac{Q\big(p^{\mathrm{f}}_{i-1}(s+1), r^{\mathrm{f}}_{i-1}(s+1), p^{\mathrm{b}}_i(s+1), r^{\mathrm{b}}_i(s+1), N_{i-1}\big)}{p} + \frac{1}{r} \right), \quad 2 \le i \le M,
\end{aligned}
\tag{8}
$$

with boundary conditions

$$
\begin{aligned}
&p^{\mathrm{f}}_1(s) = p, \quad r^{\mathrm{f}}_1(s) = r, \quad p^{\mathrm{b}}_M(s) = p,\\
&r^{\mathrm{b}}_M(s) = r, \quad s = 0, 1, 2, \ldots,
\end{aligned}
\tag{9}
$$

and initial conditions

$$
p^{\mathrm{f}}_i(0) = p, r^{\mathrm{f}}_i(0) = r, \quad i = 2, \ldots, M-1, \tag{10}
$$

where function Q is defined by Equations (6).

The recursion of Equations (8)–(10) is convergent and the following limits exist:

$$
\begin{aligned}
&\lim_{s\to\infty} p^{\mathrm{f}}_i(s) =: p^{\mathrm{f}}_i, \ \lim_{s\to\infty} p^{\mathrm{b}}_i(s) =: p^{\mathrm{b}}_i,\\
&\lim_{s\to\infty} r^{\mathrm{f}}_i(s) =: r^{\mathrm{f}}_i, \ \lim_{s\to\infty} r^{\mathrm{b}}_i(s) =: r^{\mathrm{b}}_i, \ i = 1, \ldots, M.
\end{aligned}
\tag{11}
$$

Using these limits, the estimate, PR, of the production rate is defined as follows:

$$
\begin{aligned}
&PR(p_1, r_1, p_2, r_2, \ldots, p_M, r_M, N_1, N_2, \ldots, N_{M-1})\\
&\quad = e^{\mathrm{f}}_M = e^{\mathrm{b}}_1 = e^{\mathrm{f}}_i\big[1 - Q\big(p^{\mathrm{b}}_{i+1}, r^{\mathrm{b}}_{i+1}, p^{\mathrm{f}}_i, r^{\mathrm{f}}_i, N_i\big)\big]\\
&\quad = e^{\mathrm{b}}_{i+1}\big[1 - Q\big(p^{\mathrm{f}}_i, r^{\mathrm{f}}_i, p^{\mathrm{b}}_{i+1}, r^{\mathrm{b}}_{i+1}, N_i\big)\big],
\end{aligned}
\tag{12}
$$

$$
\begin{aligned}
&\quad = e\big[1 - Q\big(p^{\mathrm{b}}_{i+1}, r^{\mathrm{b}}_{i+1}, p^{\mathrm{f}}_i, r^{\mathrm{f}}_i, N_i\big)\big]\\
&\qquad \times \big[1 - Q\big(p^{\mathrm{f}}_{i-1}, r^{\mathrm{f}}_{i-1}, p^{\mathrm{b}}_i, r^{\mathrm{b}}_i, N_{i-1}\big)\big],
\end{aligned}
$$

where

$$
e^{\mathrm{f}}_i = \frac{r^{\mathrm{f}}_i}{r^{\mathrm{f}}_i + p^{\mathrm{f}}_i}, \ e^{\mathrm{b}}_i = \frac{r^{\mathrm{b}}_i}{r^{\mathrm{b}}_i + p^{\mathrm{b}}_i}, \quad i = 2, \ldots, M-1, \tag{13}
$$

and e is defined by Equation (1).

It is shown in Chiang *et al.* (2000) that Equation (12) provides an accurate estimate of the production rate, typically within $1 - 2\%$ of its value determined by simulations.

Equations (4)–(13) are used below to evaluate k_E.

5. Lean buffering in two-machine lines

When $p_1 = p_2 = p$ and $r_1 = r_2 = r$, Equation (6) simplifies to:

$$
\begin{aligned}
Q(p_1, r_1, p_2, r_2, N_1) &= Q(p, r, N)\\
&= \frac{2p}{(p+r)[2 + (p+r)N]}.
\end{aligned}
$$

Therefore, it follows from Equation (4) that the equation that defines N_E, i.e., the lean buffer capacity necessary and sufficient to ensure line efficiency E, can be written as follows:

$$
PR = Ee = e[1 - Q(p, r, N_E)].
$$

Solving for N_E results in

$$
N_E = \begin{cases} \dfrac{2(1-e)(E-e)}{p(1-E)}, & \text{if } e < E,\\ 0, & \text{otherwise.} \end{cases} \tag{14}
$$

Since $k_E = N_E / T_{\mathrm{down}} = rN_E$, we obtain:

Proposition 1. (Enginarlar et al., 2002): *In two-machine lines defined by assumption 1–5, the lean level of buffering is given by*:

$$
k_E(M = 2) = \begin{cases} \dfrac{2e(E-e)}{1-E}, & \textit{if } e < E,\\ 0, & \textit{otherwise.} \end{cases} \tag{15}
$$

Figures 3(a) and 4(a) show the behavior of k_E as a function of e and E, respectively. From these figures and

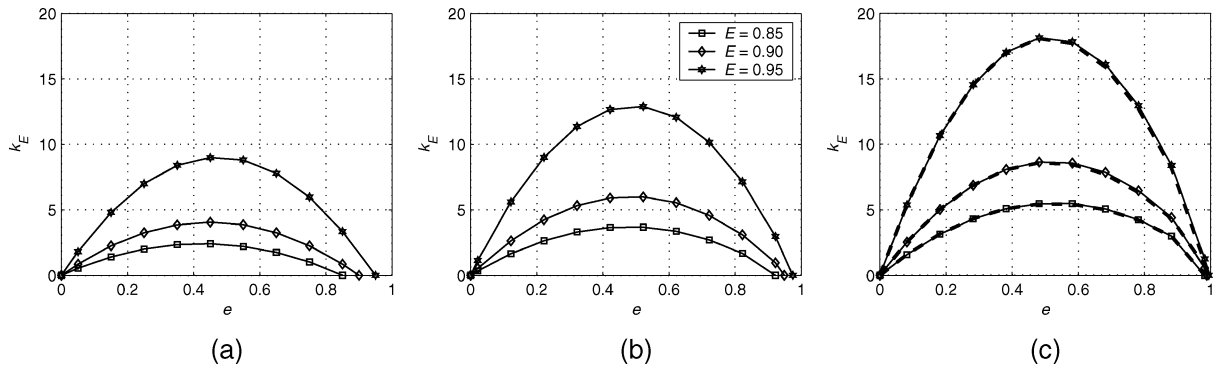

Fig. 3. Lean level of buffering as a function of machine efficiency when: (a) $M = 2$; (b) $M = 3$; and (c) $M = 10$.

Equation (15) we observe:

- LLB is a quadratic function of e, increasing for $e < 0.5E$ and decreasing for $e > 0.5E$.
- LLB does not depend explicitly on T_{up} and T_{down}.
- If $E > e$, the lean system *must* contain buffering. For instance, if $e = 0.85$ and $E = 0.95$, LLB is approximately three downtimes.
- If $\mathrm{E} \leq e$, the lean system needs no buffering. For example, if $e = 0.85$ and $E = 0.85$, $N = 0$ is acceptable.
- LLB is a monotonically increasing function of E with hyperbolically increasing rate, approaching infinity when $E \to 1$.

6. Lean buffering in three-machine lines

Since no closed formula for PR in three-machine lines is available, $k_E(M = 3)$ cannot be evaluated using arguments similar to those of Section 5. However, based on the recursive aggregation procedure of Equations (8)–(10), the following can be proved:

Proposition 2. *In three-machine serial lines defined by assumptions 1–5, the lean level of buffering is given by*:

$$k_E(M=3) = \begin{cases} \dfrac{(e + e\sqrt{E} - 2)(1+\sqrt{E})(1 - e\sqrt{E})e\sqrt{E}}{(1-\sqrt{E})[e\sqrt{E}(2 - e - e\sqrt{E}) + (1-e)(1+\sqrt{E})(1-e\sqrt{E})]} \ln\left(\dfrac{1 - e\sqrt{E}}{(1-e)(1+\sqrt{E})}\right), & \text{if } e < \sqrt{E}, \\ 0, & \text{otherwise}. \end{cases} \tag{16}$$

Proof. See the Appendix. ■

Figures 3(b) and 4(b) show the behavior of $k_E(M = 3)$ as a function of e and E, respectively. From these figures and Equation (16) we observe:

- Although $k_E(e)$ is not quadratic, its qualitative behavior is similar to a parabola, reaching its maximum around $e = 0.5\sqrt{E}$.
- LLB again is independent of T_{up} and T_{down} explicitly.
- If $\sqrt{E} > e$, a lean system must be buffered. For example, if $e = 0.85$ and $E = 0.95$, LLB is roughly six. If $e = 0.85$ and $E = 0.85$, LLB is close to one; note that for $M = 2$, this case required no buffering.
- If $\sqrt{E} \leq e$, the lean system needs no buffering. Since $0 < E < 1$ and, therefore, $\sqrt{E} > E$, the region of e's where no buffer is needed is smaller than that for $M = 2$.

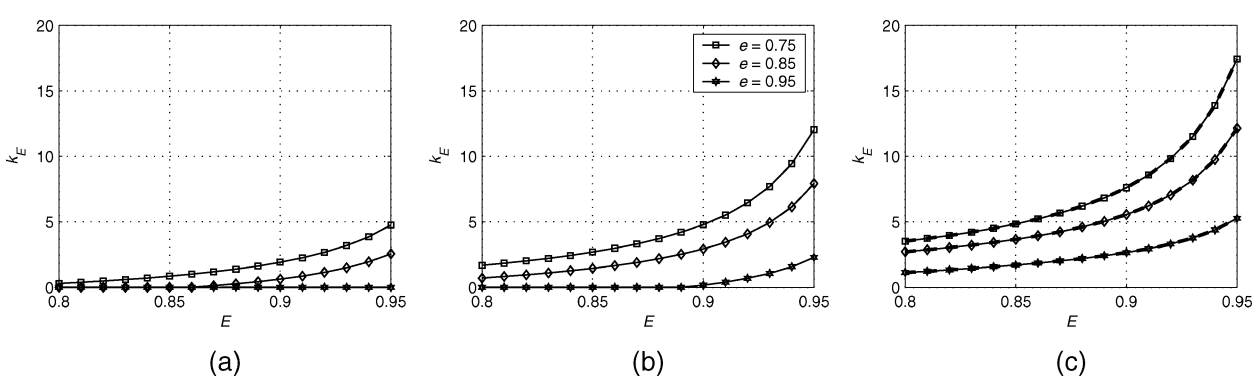

Fig. 4. Lean level of buffering as a function of line efficiency when: (a) $M = 2$; (b) $M = 3$; and (c) $M = 10$.

- *LLB* is monotonically increasing as a function of E, with an increasing rate.

7. Lean buffering in lines with more than three machines

Unlike the two- and three-machine cases, no exact *closed* formula for $k_E(M > 3)$ can be derived. However, based on the recursive aggregation procedure of Equations (8)–(10), the following can be proved:

Proposition 3. *In $M > 3$-machine serial lines defined by assumptions 1–5, the lean level of buffering is given by*:

$$k_E(M > 3) = \begin{cases} \dfrac{e(1-Q)(eQ+1-e)(eQ+2-2e)(2-Q)}{Q(2e-2eQ+eQ^2+Q-2)} \ln\left(\dfrac{E-eE+eEQ-1+e-2eQ+eQ^2+Q}{(1-e-Q+eQ)(E-1)}\right), & \text{if } e < E^{\frac{1}{M-1}}, \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

where

$$Q = Q\left(p^{\mathrm{f}}_{M-2}, r^{\mathrm{f}}_{M-2}, p^{\mathrm{b}}_{M-1}, r^{\mathrm{b}}_{M-1}, N\right), \tag{18}$$

is defined by Equations (6) and (7) and p^{f}_{M-2}, r^{f}_{M-2}, p^{b}_{M-1}, r^{b}_{M-1} are the steady states of the aggregation procedure of Equations (8)–(10).

Proof. See the Appendix. ■

It follows from Chiang *et al.* (2000), that Equation (18) is an estimate of the probability that buffer b_{M-2} is empty. This function can be evaluated using the recursive procedure of Equations (8)–(10). The behavior of Q is illustrated in Fig. 5(a–c) as a function of e for several values of E and $M = 10$ (solid lines). This behavior implies that Q, as a function of e, can be approximated by the expression:

$$\hat{Q} = a + b\exp\left\{-\frac{\alpha - e}{\beta}\right\}, \tag{19}$$

where a, b, α and β are functions of E and M to be selected so that the approximation is sufficiently precise. We derive this approximation as follows:

For $M = 3$ and any e, function Q can be calculated in the closed form as (see the proof of Proposition 2 in the Appendix):

$$Q = 1 - \sqrt{E}. \tag{20}$$

For any M and $e = E^{\frac{1}{M-1}}$ (i.e., when no buffering is required), function Q again can be easily calculated as:

$$Q = 1 - E^{\frac{M-2}{M-1}}. \tag{21}$$

Finally, if we assume that $k_E(e)$ is a parabolic function reaching its maximum at:

$$e = \frac{1}{2}E^{\frac{1}{M-1}},$$

then function Q can be evaluated as:

$$Q \approx 1 - E^{\frac{1}{2}[1+(\frac{M-3}{M-2})M/4]}. \tag{22}$$

Using these expressions, we select a, b and α of Equation (19) as follows:

$$\begin{aligned} a &= 1 - E^{\frac{1}{2}[1+(\frac{M-3}{M-1})^{M/4}]}, \\ b &= E^{\frac{1}{2}[1+(\frac{M-3}{M-2})^{M/4}]} - E^{\frac{M-3}{M-1}}, \\ \alpha &= E^{\frac{1}{M-1}}. \end{aligned} \tag{23}$$

To select β, we observe that smaller values of E require larger time constants of the exponential behavior of Q as a function of e. By trial and error, we found that this can be captured by selecting β as:

$$\beta = 1 - E. \tag{24}$$

Thus, we obtain the following approximation of Q:

$$\hat{Q} = 1 - E^{\frac{1}{2}[1+(\frac{M-3}{M-1})^{M/4}]} + \left(E^{\frac{1}{2}[1+(\frac{M-3}{M-1})^{M/4}]} - E^{\frac{M-2}{M-1}}\right) \times \exp\left\{-\left(\frac{E^{\frac{1}{M-1}} - e}{1-E}\right)\right\}. \tag{25}$$

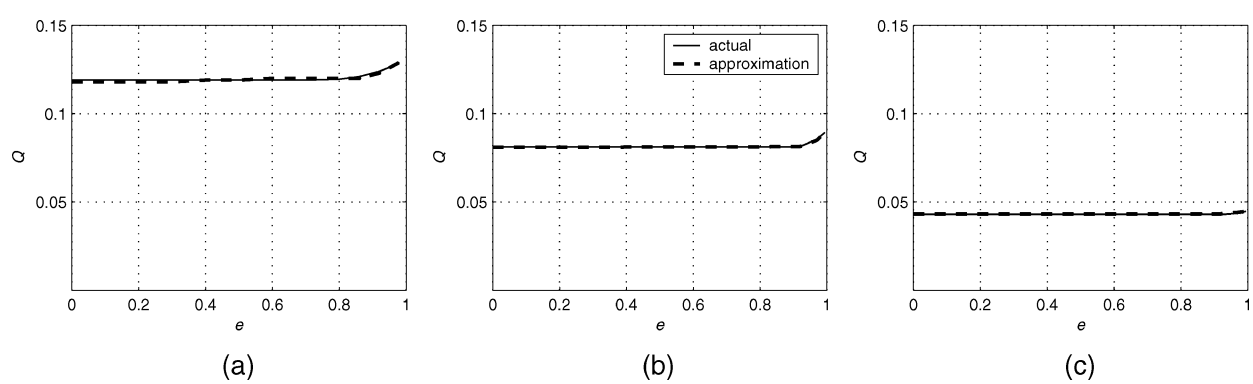

Fig. 5. The behavior of functions Q and $\hat{Q}$ against e when: (a) $M = 10$, $E = 0.85$; (b) $M = 10$, $E = 0.9$; and (c) $M = 10$, $E = 0.95$.

Table 1. Accuracy of approximating Q by $\hat{Q}$ (percent of error) when $M = 5$

e	$E = 0.85$	$E = 0.9$	$E = 0.95$
0.1	3.38	3.28	3.13
0.2	3.23	3.15	3.06
0.3	3.03	3.01	2.99
0.4	2.76	2.77	2.83
0.5	2.46	2.59	2.79
0.6	2.27	2.28	2.49
0.7	2.10	1.96	2.14
0.8	1.85	1.81	1.36
0.9	3.53	2.89	0.95

Table 2. Accuracy of approximating Q by $\hat{Q}$ (percent of error) when $M = 10$

e	$E = 0.85$	$E = 0.9$	$E = 0.95$
0.1	0.75	0.48	0.24
0.2	0.52	0.32	0.16
0.3	0.29	0.12	0.04
0.4	0.07	0.11	−0.11
0.5	−0.25	−0.40	−0.26
0.6	−0.19	−0.69	−0.58
0.7	−0.35	0.87	−1.07
0.8	0.76	0.60	−1.83
0.9	2.72	1.16	−2.81

Note that for $M = 3$ this expression coincides with Equation (20) whereas for $e = E^{\frac{1}{M-1}}$ it coincides with Equation (21).

The behavior of $\hat{Q}$ is shown in Fig. 5(a–c) by broken lines. Quantitatively, the accuracy of this approximation (i.e., $(Q - \hat{Q})/Q) \times 100\%$) is characterized in Tables 1–3. As one can see, it is within 4%. Since this is commensurable with the accuracy of the data available in most practical applications, we conclude that Equations (25) and (17) where $\hat{Q}$ is substituted for Q, i.e.:

$$\hat{k}_E(M > 3) = \begin{cases} \dfrac{e(1-\hat{Q})(e\hat{Q}+1-e)(e\hat{Q}+2-2e)(2-\hat{Q})}{\hat{Q}(2e-2e\hat{Q}+e\widehat{Q^2}+\hat{Q}-2)} \ln\left(\dfrac{E-eE+eE\hat{Q}-1+e-2e\hat{Q}+e\widehat{Q^2}+\hat{Q}}{(1-e-\hat{Q}+e\hat{Q})(E-1)}\right), & \text{if } e < E^{\frac{1}{M-1}}, \\ 0, & \text{otherwise}, \end{cases} \tag{26}$$

provide a closed, however approximate, formula for LLB in $M >$ three-machine case. Note that for $M = 3$, Equations (25) and (26) coincide with Equation (16).

Equations (25) and (26) define $\hat{k}_E$ as a function of three variables: e, E and M. For a fixed M, this function is qualitatively similar to those for $M = 2$ and 3. The solid lines in Figs. 3(c) and 4(c) show this function for $M = 10$; the dashed lines, which almost overlay with the solid lines, show the values of k_E (calculated according to the recursive procedure of Equations (8)–(10) and Equations (17) and (18)). Obviously, LLB increased significantly and the area of no buffering decreased in comparison with the $M = 3$ case. For instance, if $E = 0.9$, the smallest e that ensures this E without buffers is 0.9884.

Equations (25) and (26) can be used to analyze the behavior of $\hat{k}_E$ as a function of M. This behavior is illustrated in Fig. 6(a–c) where the values of k_E, calculated according to Equations (8)–(10) and Equations (17) and (18) are also shown (broken lines). The errors, i.e. $((\hat{k}_E - k_E)/k_E) \times 100\%$, are quantified in Tables 4–6. From these tables and Fig. 6(a–c) we conclude:

- Equations (25) and (26) provide a characterization of LLB typically within a 5% error; in a few cases, however, the errors may reach 10%.
- LLB is a monotonically increasing function of M.
- However, the rate of increase is decreasing, leveling off at about $M = 15$. Thus, the LLB necessary for $M = 15$ is practically acceptable for any $M > 15$.

Remark: It should be pointed out that while for any given system the recursive procedure of Equations (8)–(10) can be used to evaluate the required lean buffering, this

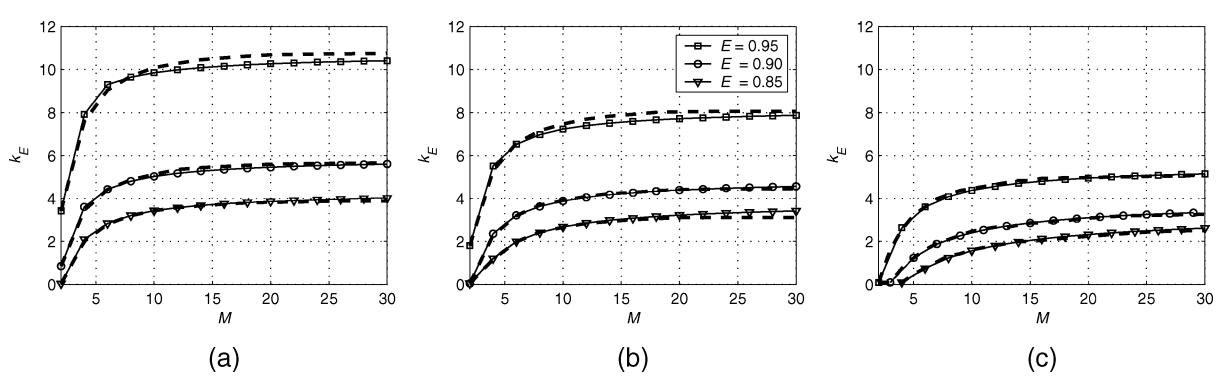

Fig. 6. Lean level of buffering as a function of M, approximated (solid lines) and exact (dashed lines) when: (a) $e = 0.85$; (b) $e = 0.9$; and (c) $e = 0.95$.

Table 3. Accuracy of approximating Q by $\hat{Q}$ (percent of error) when $M = 20$

e	$E = 0.85$	$E = 0.9$	$E = 0.95$
0.1	−0.70	−0.38	−0.09
0.2	−0.43	−0.19	−0.01
0.3	−0.13	−0.05	−0.11
0.4	0.16	0.66	0.29
0.5	0.42	0.99	0.88
0.6	0.31	1.15	1.40
0.7	0.41	0.45	0.46
0.8	0.92	0.56	1.95
0.9	3.21	1.03	1.96

Table 4. Accuracy of approximating k_E by $\hat{k}_E$ (percent of error) when $e = 0.85$

M	$E = 0.85$	$E = 0.9$	$E = 0.95$
5	−2.05	−1.47	−5.03
10	0.05	−1.86	2.36
15	−0.22	2.69	3.65
20	−1.12	2.45	3.87
25	−2.24	1.65	3.51
30	−3.34	1.01	3.25

Table 5. Accuracy of approximating k_E by $\hat{k}_E$ (percent of error) when $e = 0.9$

M	$E = 0.85$	$E = 0.9$	$E = 0.95$
5	−0.54	−1.37	−1.17
10	−0.73	0.76	3.23
15	−2.18	1.21	4.05
20	−4.61	−0.03	3.77
25	−7.67	−1.41	2.82
30	−9.81	−2.59	2.23

Table 6. Accuracy of approximating k_E by $\hat{k}_E$ (percent of error) when $e = 0.95$

M	$E = 0.85$	$E = 0.9$	$E = 0.95$
5	−2.08	0.18	2.96
10	−3.16	−2.52	2.01
15	−2.54	−0.32	1.71
20	−3.47	−0.52	−0.12
25	−2.71	−1.73	−1.18
30	−4.22	−2.92	−1.24

Table 7. Rule-of-thumb for selecting a lean level of buffering in serial production lines with 15 or more exponential machines

e	$E = 0.85$	$E = 0.90$	$E = 0.95$
0.85	3.8	5.5	10.5
0.90	3.0	4.5	8
0.95	2.5	3.2	4.5

approach does not provide information on *LLB* as a function of system parameters, which is given in the closed analytical expressions of Equations (15) and (16) and Equations (25) and (26).

8. Discussion and extension

8.1. *Does a large LLB imply a large buffer?*

It follows from Section 7, that the *LLB* for $M \geq 15$-machine lines may be selected according to the "rule-of-thumb" specified in Table 7. Some of the entries of this table are quite large. A question arises: are buffers with a large *LLB*, say $k_E = 10$, large or small? This question is ill-posed. Indeed, $k_E = 10$ may represent either a large or a small buffer, depending on the value of T_{down}: if T_{down} is, say, 100 cycle times, then the buffer capacity, N_E, is 1000, and therefore, it may be considered as large; if $T_{\text{down}} = 1$, then $N_E = 10$, and the buffer is small.

Equally ill-posed is the question: is a buffer of capacity $N_E = 100$ large or small? If $T_{\text{down}} = 100$, then $k_E = 1$, and the buffer may be viewed as very small, and moreover, insufficient if $e \leq 0.95$ and $E \geq 0.85$ (see Table 7); if $T_{\text{down}} = 2$, then $k_E = 50$, and the buffer is an order of magnitude larger than required for any reasonable line efficiency.

8.2. *Lean buffering for non-exponential machines*

Consider a serial production line defined by assumptions 1–5 with the only difference that the up- and downtimes of the machines are distributed arbitrarily, rather than exponentially. Assume that the coefficients of variation of these distributions (i.e., the ratio between the standard deviation and the expected value) are less than one. This assumption is supported by empirical evidence (Inman, 1999). What is the relationship between *LLB* for such a line (denoted as $k_E^{\text{non-exp}}$) with that calculated in Sections 5–7 (i.e., k_E)? Although no rigorous answer to this question is available, based on numerous simulations (see also Enginarlar *et al.* (2002)), we formulate the following:

Conjecture 1. If the coefficients of variation of up- and downtimes are less than one, then k_E is the upper bound of $k_E^{\text{non-exp}}$, i.e.:

$$k_E^{\text{non-exp}} < k_E.$$

In the framework of this conjecture, results of this paper shed light on lean buffering in serial lines with identical non-exponential machines. A justification of this conjecture is a part of future work.

9. Conclusions

This paper provides a contribution to both the theory and practice of production systems engineering. From the theory point of view, it offers an analytical characterization

of the level of buffering necessary and sufficient to ensure a desired production rate in a simple model of a manufacturing system. From a practical point of view, it offers production systems designers a method for selecting an initial level of buffering, which might be used as an initial condition for subsequent analysis and simulations. Finally, the insights into lean buffering described in this work (for instance, the rule-of-thumb of Table 7) may be useful for factory floor personnel as a guide for improving production systems leanness.

Acknowledgement

This work was supported by NSF grant DMI-0245377. The authors are grateful to three referees for their helpful comments.

References

Altiok, T. and Stidham, S.S. (1983) The allocation of interstage buffer capacities in production lines. *IIE Transactions*, **15**, 292–299.

Buzacott, J.A. (1967) Automatic transfer lines with buffer stocks. *International Journal of Production Research*, **5**, 183–200.

Buzacott, J.A. and Shantikumar, J.G. (1993) *Stochastic Models of Manufacturing Systems*, Prentice Hall, Englewood Cliffs, NJ.

Caramanis, M. (1987) Production line design: a discrete event dynamic system and generalized Benders decomposition approach. *International Journal of Production Research*, **25**, 1223–1234.

Chiang, S.-Y., Kuo, C.-T. and Meerkov, S.M. (2000) DT-bottlenecks in serial production lines: theory and application. *IEEE Transactions on Robotics and Automation*, **16**, 567– 580.

Chow, W.-M. (1987) Buffer capacity analysis for sequential production lines with variable processing times. *International Journal of Production Research*, **25**, 1183–1196.

Enginarlar, E., Li, J., Meerkov, S.M. and Zhang, R.Q. (2002) Buffer capacity for accommodating machine downtime in serial production lines. *International Journal of Production Research*, **40**, 601–624.

Freeman, M.C. (1964) The effect of breakdowns and interstage storage on production capacity. *Journal of Industrial Engineering*, **15**, 194–200.

Gershwin, S.B. (1994) *Manufacturing Systems Engineering*, Prentice Hall, Englewood Cliffs, NJ.

Gershwin, S.B. and Goldis, Y. (1995) Efficient algorithms for transfer line design. Report LMP-95-005, MIT Laboratory for Manufacturing and Productivity, MIT, Cambridge, MA.

Gershwin, S.B. and Schor, J.E. (2000) Efficient algorithms for buffer space allocation. *Annals of Operations Research*, **93**, 117–144.

Goldratt, E.M. and Cox, J. (1986) *The Goal*, North River Press, Croton-on-Hudson, NY.

Harris, J.H. and Powell, S.G. (1999) An algorithm for optimal buffer placement in reliable serial lines. *IIE Transactions*, **31**, 287–302.

Hillier, F. and So, K.C. (1991a) The effect of the coefficient of variation of operation times on the allocation of storage space in production line systems. *IIE Transactions*, **23**, 198–206.

Hillier, F.S. and So, K.C. (1991b) The effect of machine breakdowns and internal storage on the performance of production line systems. *International Journal of Production Research*, **29**, 2043–2055.

Inman, R.R. (1999) Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems. *Production and Operation Management*, **4**, 409–432.

Jacobs, D. and Meerkov, S.M. (1995) A system-theoretic property of serial production lines: improvability. *International Journal of Systems Science*, **26**, 755–785.

Jafari, M.A. and Shanthikumar, J.G. (1989) Determination of optimal buffer storage capacity and optimal allocation in multistage automatic transfer lines. *IIE Transactions*, **21**, 130–134.

Liu, C.-M. and Lin, C.-L. (1994) Performance evaluation of unbalanced serial production lines. *International Journal of Production Research*, **32**, 2897–2914.

McClain, J.O., Conway, R., Maxwell, W. and Thomas, L.J. (1988) The role of work-in-process inventory in serial production lines. *Operations Research*, **36**, 229–241.

Park, T. (1993) A two-phase heuristic algorithm for determining buffer sizes in production lines. *International Journal of Production Research*, **31**, 613–631.

Plambeck, E.L., Fu, B.-R., Robinson, S.M. and Suri, R. (1993) Throughput optimization in tandem production lines via nonsmooth programming, *Proceedings of the 1993 Summer Computer Simulation Conference*, San Diego, CA, 70–75.

Powell, S.G. (1994) Buffer allocation in unbalanced three-station lines. *International Journal of Production Research*, **32**, 2201–2217.

Powell, S.G. and Pyke, D.F. (1996) Allocation of buffers to serial production lines with bottlenecks. *IIE Transactions*, **28**, 18–29.

Seong, D., Change, S.Y. and Hong, Y. (1995) Heuristic algorithm for buffer allocation in a production line with unreliable machines. *International Journal of Production Research*, **33**, 1989–2005.

Sheskin, T.J. (1976) Allocation of interstage storage along an automatic production line. *AIIE Transactions*, **8**, 146–152.

Smith, J.M. and Daskalaki, S. (1988) Buffer space allocation in automated assembly lines. *Operations Research*, **36**, 343–357.

Soyster, A.L., Schmidt, J.W. and Rohrer, M.W. (1979) Allocation of buffer capacities for a class of fixed cycle production lines. *AIIE Transactions*, **11**, 140–146.

Tempelmeier, H. (2003) Practical considerations in the optimization of flow production systems. *International Journal of Production Research*, **41**, 149–170.

Yamashita, H. and Altiok, T. (1998) Buffer capacity allocation for a desired throughput of production lines. *IIE Transactions*, **30**, 883–891.

Yamashita, H. and Suzuki, S. (1987) An approximate solution method for optimal buffer allocation in serial n-stage automatic production lines. *Transactions of the Japanese Society of Mechanical Engineers*, **57**, 807–817.

Appendix

It is convenient first to prove Proposition 3 and then specialize it to Proposition 2.

Proof of Proposition 3. The steady states of the recursive procedure of Equations (8)–(10) are defined by:

$$\begin{aligned} p_i^{\mathrm{b}} &= \frac{p}{1 - Q\big(p_{i+1}^{\mathrm{b}}, r_{i+1}^{\mathrm{b}}, p_i^{\mathrm{f}}, r_i^{\mathrm{f}}, N_i\big)}, \quad 1 \le i \le M-1, \\ r_i^{\mathrm{b}} &= \frac{pr}{Q\big(p_{i+1}^{\mathrm{b}}, r_{i+1}^{\mathrm{b}}, p_i^{\mathrm{f}}, r_i^{\mathrm{f}}, N_i\big)r + p}, \quad 1 \le i \le M-1, \quad \text{(A1)} \\ p_i^{\mathrm{f}} &= \frac{p}{1 - Q\big(p_{i-1}^{\mathrm{f}}, r_{i-1}^{\mathrm{f}}, p_i^{\mathrm{b}}, r_i^{\mathrm{b}}, N_{i-1}\big)}, \quad 2 \le i \le M, \\ r_i^{\mathrm{f}} &= \frac{pr}{Q\big(p_{i-1}^{\mathrm{f}}, r_{i-1}^{\mathrm{b}}, p_i^{\mathrm{b}}, r_i^{\mathrm{b}}, N_{i-1}\big)r + p}, \quad 2 \le i \le M, \end{aligned}$$

where function Q is given by Equations (6) and (7).

When N is selected so that the efficiency of the production line is E, it follows from Equation (12) that:

$$Ee = e\big[1 - Q\big(p_{M-1}^{\mathrm{f}}, r_{M-1}^{\mathrm{f}}, p, r, N_E\big)\big],$$

which implies that

$$Q\big(p^{\mathrm{f}}_{M-1}, r^{\mathrm{f}}_{M-1}, p, r, N_E\big) = 1 - E.$$

This, along with Equation (6), provides the equation for N_E:

$$\frac{\big(1 - e^{\mathrm{f}}_{M-1}\big)(1-\phi)}{1 - \phi \exp\{-\beta N_E\}} = 1 - E,$$

i.e.,

$$N_E = -\frac{1}{\beta} \ln\left(\frac{E - \phi - e^{\mathrm{f}}_{M-1} + e^{\mathrm{f}}_{M-1}\phi}{\phi(E-1)}\right).$$

Therefore, the lean level of buffering, $k_E = N_E / T_{\mathrm{down}}$, is given by

$$k_E = -\frac{r}{\beta} \ln\left(\frac{E - \phi - e^{\mathrm{f}}_{M-1} + e^{\mathrm{f}}_{M-1}\phi}{\phi(E-1)}\right). \tag{A2}$$

In these expressions, as it follows from Equations (7), (12) and (A1):

$$e^{\mathrm{f}}_{M-1} = e(1-Q),$$
$$\phi = \frac{(1-Q)(1-e)}{1-e+eQ},$$
$$\beta = \frac{\big(r^{\mathrm{f}}_{M-1} + r + p^{\mathrm{f}}_{M-1} + p\big)\big(p^{\mathrm{f}}_{M-1} r - p r^{\mathrm{f}}_{M-1}\big)}{\big(r^{\mathrm{f}}_{M-1} + r\big)\big(p^{\mathrm{f}}_{M-1} + p\big)},$$

where

$$p^{\mathrm{f}}_{M-1} = \frac{p}{1+Q}, r^{\mathrm{f}}_{M-1} = \frac{pr}{Qr+p},$$
$$Q = Q\big(p^{\mathrm{f}}_{M-2}, r^{\mathrm{f}}_{M-2}, p^{\mathrm{b}}_{M-1}, r^{\mathrm{b}}_{M-1}, N_E\big). \tag{A3}$$

Using these expressions and the identities

$$p = \frac{r(1-e)}{e},$$
$$p + r = \frac{r}{e},$$

Equation (A2) can be re-written as:

$$k_E = \frac{e(1-Q)(eQ+1-e)(eQ+2-2e)(2-Q)}{Q(2e-2eQ+eQ^2+Q-2)} \times \ln\left(\frac{E - eE + eEQ - 1 + e - 2eQ + eQ^2 + Q}{(1-e-Q+eQ)(E-1)}\right). \tag{A4}$$

This proves the upper line of Equation (17).

To prove the bottom line, we observe that if $N = 0$, the recursive procedure of Equations (8)–(10) results in

$$PR = e^M = eE.$$

This implies that if $e \geq E^{\frac{1}{M-1}}$, $M \geq 2$, line efficiency E is attained with $k_E = 0$.

Proof of Proposition 2. In the case of $M = 3$, function Q defined by Equation (A3) can be evaluated explicitly. Indeed, for $M = 3$, Equation (A3) can be written as follows:

$$Q = Q\big(p, r, p^{\mathrm{b}}_2, r^{\mathrm{b}}_2, N_E\big).$$

To evaluate this function, we observe that, as it follows from Equation (12):

$$Q\big(p, r, p^{\mathrm{b}}_2, r^{\mathrm{b}}_2, N_E\big) = Q\big(p, r, p^{\mathrm{f}}_2, r^{\mathrm{f}}_2, N_E\big)$$

and, moreover,

$$Ee = e(1-Q)^2.$$

This implies that:

$$Q = 1 - \sqrt{E}.$$

Substituting this expression into Equation (A4) results in Equation (16).

Biographies

Emre Enginarlar is a Postdoctoral Research Associate at Los Alamos National Laboratory. He received his B.S. in EE from the Middle East Technical University in Ankara, Turkey, M.S. and Ph.D. in EE: Systems Science and Engineering from the University of Michigan, Ann Arbor in 1997, 1999, and 2003, respectively. His research interests focus on modeling, analysis, design and continuous improvement of manufacturing systems.

Jingshan Li is a Staff Research Engineer at Manufacturing Systems Research Lab, General Motors Research and Development Center, Warren, MI. He received his bachelor's degree from the Department of Automation, Tsinghua University, Beijing, China, in 1989 and a Master's degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China in 1992. In 2000, he received his Ph.D. in Electrical Engineering—Systems from the University of Michigan, Ann Arbor, MI. His primary research interests are in the modeling and analysis of complex manufacturing systems. He is a member of IIE.

Semyon M. Meerkov received his MSEE degree from the Polytechnic of Kharkov, Ukraine, in 1962 and Ph.D. 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 of the University of Michigan, Ann Arbor, MI. He has held visiting positions at UCLA (1978-1979), Stanford (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 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.

Contributed by the Engineering Design Department