LEAN BUFFERING IN SERIAL PRODUCTION LINES WITH BERNOULLI MACHINES

A. B. HU AND S. M. MEERKOV

Lean buffering is the smallest buffer capacity necessary to ensure the desired production rate of a manufacturing system. In this paper, analytical methods for selecting lean buffering in serial production lines are developed under the assumption that the machines obey the Bernoulli reliability model. Both closed-form expressions and recursive approaches are investigated. The cases of identical and nonidentical machines are analyzed. Results obtained can be useful for production line designers and production managers to maintain the required production rate with the smallest possible inventories.

1. Introduction

Production lines with unreliable machines usually contain finite capacity buffers intended to attenuate mutual perturbations of the machines due to breakdowns. It is well known that the capacity of the buffers should be as small as possible, that is, lean. But how lean can lean be? This question was addressed in [2] where a method for selecting lean buffering in production systems with identical machines obeying the exponential reliability model has been developed. A generalization for machines with reliability models described by other continuous random variables (such as Weibull, gamma, and log-normal) has been reported in [3]. However, the issue of lean buffering with machine reliability models described by discrete random variables has not been addressed. In this paper, we analyze lean buffering in serial lines with Bernoulli machines and, along with identical machines, consider the case of nonidentical ones as well.

The reason for considering Bernoulli machines is that in many assembly operations the downtime is short and comparable with the cycle time. This happens because the stoppage of an assembly line is often due to a pallet being jammed on a conveyor or due to the desire of an operator to complete the operation with the highest possible quality. In these cases, the Bernoulli reliability model [7], according to which the probability that a

machine produces a part during a cycle time is p and that a machine fails to do so is $1 - p$, is more appropriate than the exponential one where up- and downtimes of the machine are distributed exponentially.

The reason for considering nonidentical machines is due to the fact that in most production systems the machines cannot be viewed as identical, as far as their efficiency is concerned. Indeed, even if originally the line was designed having machines with identical performance, in the course of time their characteristics drift (due, e.g., to wear and/or various adjustments during maintenance), which results in nonidentical performance. Thus, considering production lines with nonidentical Bernoulli machines is of practical importance. This paper is intended to contribute to this end.

To place the current paper in the framework of existing literature, we remark that the problem of buffer capacity allocation in serial lines has been studied quantitatively for over 50 years, and a large number of publications are 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 [1, 4–6, 9–11], respectively. In these publications, machine reliability models are characterized by continuous random variables, in most cases, exponential ones. The current work contributes to this literature by analyzing a discrete random variable case.

The outline of this paper is as follows. Section 2 below presents the problem formulation. In Section 3, the approach is described. Sections 4 and 5 present the main results— for identical and nonidentical machines, respectively. Finally, in Section 6 the conclusions are formulated. All proofs are given in the appendix.

2. Problem formulation

Consider a serial production line shown in Figure 2.1, 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) The machines have identical cycle time T_c. The time axis is slotted with the slot duration T_c. Machines begin operating at the beginning of each time slot.
(iii) Each buffer is characterized by its capacity, $N_i < \infty, 1 \le i \le M - 1$.
(iv) Machine i is starved during a time slot if buffer $i - 1$ is empty at the beginning of the time slot. Machine 1 is never starved for parts.
 (v) Machine i is blocked during a time slot if buffer i has N_i parts at the beginning of the time slot, and machine $i + 1$ fails to take a part during the time slot. Machine M is never blocked by ready goods buffer.
(vi) Machines obey the Bernoulli reliability model, that is, machine $i, i = 1,\ldots,M$, being neither blocked nor starved during a time slot, produces a part with probability p_i and fails to do so with probability $1 - p_i$. Parameter p_i is referred to as the *efficiency* of machine i.

Let PR denote the production rate of this line, that is, the average number of parts produced by the last machine in the steady state. Let PR_∞ be the production rate of the line
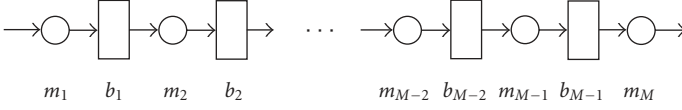
Figure 2.1. Serial production line.

consisting of the same machines but having buffers with infinite capacity (and, therefore, having the largest possible production rate; see [7]). Following [2], introduce the notion of *line efficiency*:

$$E = \frac{\text{PR}}{\text{PR}_\infty}.$$
(2.1)

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

$$N_{1,E}, \ldots, N_{M-1,E}$$
(2.2)

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

The problem addressed in this paper is *to develop analytical methods for calculating* LBC *as a function of machine efficiency* p_i, $i = 1, \ldots, M$, *line efficiency* E, *and the number of machines in the system* M. The case of identical machines, that is, $p_i =: p$, $i = 1, \ldots, M$, is carried out in Section 4 while the case of nonidentical machines, $p_i \neq p_j, i, j, = 1, \ldots, M$, is addressed in Section 5.

3. Approach

The approach of this paper is based on a method for performance analysis of serial lines with Bernoulli machines developed in [7]. To make this paper self-contained, it is briefly reviewed below.

For $M = 2$, using Markov chain analysis, the production rate, PR, of a serial line defined by assumptions (i)–(vi) has been shown to be

$$\text{PR} = p_1[1 - Q(p_2, p_1, N)] = p_2[1 - Q(p_1, p_2, N)],$$
(3.1)

where

$$Q(x, y, N) = \begin{cases} \dfrac{(1-x)(1-\alpha)}{1 - (x/y)\alpha^N}, & x \neq y, \\ \dfrac{1-x}{N+1-x}, & x = y, \end{cases}$$
(3.2)

$$\alpha = \frac{x(1-y)}{y(1-x)}.$$
(3.3)

For $M > 2$, no closed formulas for PR can be derived, and, therefore, a recursive aggregation procedure, which leads to an estimate of PR with accuracy typically within 1%, has been developed. This procedure consists of two parts—the backward and forward aggregations. Within the backward aggregation, the last two machines are aggregated into a single machine, using expressions (3.1). 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 efficiencies of the aggregated machines are denoted as p_i^b, $i = 1, \ldots, M − 1$, 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 efficiencies of these aggregated machines are denoted as p_i^f, $i = 2, \ldots, M$, where f stands for "forward." Then the procedure is repeated again, alternating between the backward and forward aggregations. Formally, this process is represented as

$$p_i^b(s+1) = p_i\left[1 − Q\left(p_{i+1}^b(s+1), p_i^f(s), N\right)\right], \quad i = 1, \ldots, M − 1,$$

$$p_i^f(s+1) = p_i[1 − Q(p_{i-1}^f(s+1), p_i^b(s), N)], \quad i = 2, \ldots, M, \ s = 1, 2, \ldots,$$

$$(3.4)$$

with function Q defined in (3.2) and the initial and boundary conditions given by

$$p_i^f(0) = p_i, \quad i = 1, \ldots, M, \tag{3.5}$$

$$p_1^f(s) = p_1, \qquad p_M^b(s) = p_M, \quad s = 1, 2, \ldots. \tag{3.6}$$

It has been shown in [7] that this recursive procedure is convergent and the following limits exist:

$$\lim_{s \to \infty} p_i^f(s) =: p_i^f, \quad \lim_{s \to \infty} p_i^b(s) =: p_i^b, \quad i = 1, \ldots, M. \tag{3.7}$$

Moreover,

$$p_M^f = p_1^b. \tag{3.8}$$

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

$$\text{PR} = p_M^f = p_1^b = p_{i+1}^b\left[1 − Q\left(p_i^f, p_{i+1}^b, N_i\right)\right] = p_i^f\left[1 − Q\left(p_{i+1}^b, p_i^f, N_i\right)\right], \quad i = 2, \ldots, M − 1.$$
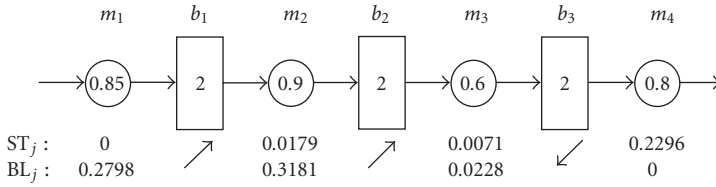
$$(3.9)$$

Figure 3.1. Bottleneck identification rules.

Expressions (3.1)–(3.9) are used throughout this paper for evaluating lean buffering of serial lines with Bernoulli machines.

In addition to the above, in Section 5 of this paper we use a method for bottleneck machine (BN) identification developed in [8], where the BN is defined as the machine i, such that

$$\frac{\partial \mathrm{PR}\left(p_1,\ldots,p_M,N_1,\ldots,N_{M-1}\right)}{\partial p_i} > \frac{\partial \mathrm{PR}\left(p_1,\ldots,p_M,N_1,\ldots,N_{M-1}\right)}{\partial p_j}, \quad \forall j \neq i. \quad (3.10)$$

To outline this method, we note that, as it has been shown in [7], aggregation procedure (3.4)–(3.6) can be used to evaluate the probabilities of blockage, BL_i, $i = 1,\ldots,M-1$, and starvation, ST_i, $i = 2,\ldots,M$, of all machines in the system. Specifically, the estimates of these probabilities are

$$\mathrm{BL}_i = p_i Q(p_{i+1}^b, p_i^f, N_i), \quad i = 1,\ldots,M-1,$$

$$\mathrm{ST}_i = p_i Q(p_{i-1}^f, p_i^b, N_{i-1}), \quad i = 2,\ldots,M,$$

$$(3.11)$$

where p_i^b, p_i^f, and Q are defined by (3.7) and (3.2), (3.3), respectively. Using these probabilities, the bottleneck machine is identified as follows.

Consider a production line and place the probabilities of starvations and blockages under each machine as shown in Figure 3.1. Assign the arrows directed from one machine to another according to the following rule. If $\mathrm{BL}_i > \mathrm{ST}_{i+1}$, the arrow is pointing from m_i to m_{i+1}. If $\mathrm{BL}_i < \mathrm{ST}_{i+1}$, the arrow is pointing from m_{i+1} to m_i. It is shown in [8] that if there is a unique machine with no emanating arrows, it is the BN in the sense of (3.10). Thus, m_3 is the BN of the production line of Figure 3.1. 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

$$S_i = \left(\mathrm{BL}_{i-1} + \mathrm{ST}_{i+1}\right) - \left(\mathrm{BL}_i + \mathrm{ST}_i\right), \quad i = 2,\ldots,M-1,$$

$$S_1 = \mathrm{ST}_2 - \mathrm{BL}_1, \qquad S_M = \mathrm{BL}_{M-1} - \mathrm{ST}_M. \quad (3.12)$$

4. Lean buffering in serial lines with identical Bernoulli machines

In this section, we assume that all machines have identical efficiency,

$$p_i =: p, \quad i = 1,\ldots,M, \tag{4.1}$$

and, in addition, all buffers are of identical capacity,

$$N_i =: N, \quad i = 1,\ldots,M - 1. \tag{4.2}$$

Assumption (4.2) is introduced in order to obtain a compact representation of the results. It should be pointed out that, as it is well known [7], a more efficient buffer allocation in systems satisfying (4.1) is the so-called inverted bawl pattern, whereby buffers in the middle of the line are larger than those at the edges. However, this leads to just a small improvement of the production rate in comparison with the uniform allocation (4.2) (typically within 1%) and, therefore, is not considered here.

4.1. Two-machine lines. In the case of two identical machines, function Q becomes

$$Q(p_1, p_2, N) = Q(p, N) = \frac{1 - p}{N + 1 - p}. \tag{4.3}$$

Therefore, using (2.1) and (3.1), we obtain the following equation for lean buffer capacity, N_E, which ensures line efficiency E:

$$E = 1 - \frac{1 - p}{N_E + 1 - p}. \tag{4.4}$$

Solving for N_E and taking into account that N_E is an integer, we obtain the following.

PROPOSITION 4.1. *The lean buffer capacity (LBC) in serial lines defined by assumptions* (i)–(vi) *with $M = 2$ and $p_1 = p_2 = p$ is given by*

$$N_E(M = 2) = \left\lceil \frac{E(1 - p)}{1 - E} \right\rceil. \tag{4.5}$$

Here and throughout this paper, $\lceil x \rceil$ denotes the smallest integer larger than x.

Note that according to (4.5), N_E cannot be less than 1. Buffering $N_E = 1$ implies that the machine itself stores a part being processed and no additional buffering between the machines is required. This can be interpreted as just-in-time (JIT) operation.

Figures 4.1(a) and 4.2(a) illustrate the behavior of the lean buffer capacity as a function of machine efficiency p and line efficiency E, respectively. From these figures and expression (4.5), we observe the following.

(α) LBC is a monotonically decreasing function of p, with a practically constant slope.

(β) LBC is a monotonically increasing function of E, exhibiting a hyperbolic behavior in $1 - E$.

(a) $M = 2$

(b) $M = 3$

(c) $M = 10$
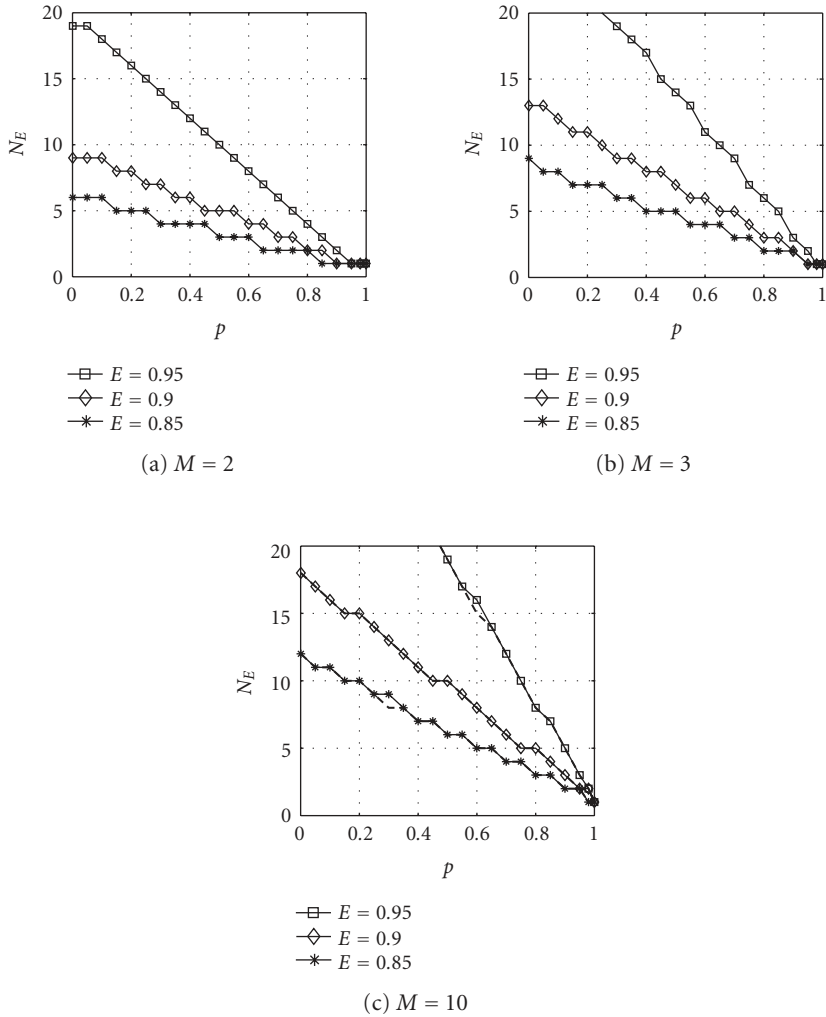
Figure 4.1. Lean buffering as a function of machine efficiency.

(γ) JIT operation is acceptable only if p's are sufficiently large. For instance, if the desired line efficiency is 0.85, JIT can be used only if $p > 0.83$, while for $E = 0.95$, p must be larger than 0.95.

(δ) In a practical range of p's, for example, $0.6 < p < 0.98$, relatively small buffers are required to achieve a large E. For instance, $N_{0.95} = 6$ if $p = 0.7$; if $p = 0.9$, $N_{0.95} = 2$.

4.2. Three-machine lines. For a three-machine line, using the aggregation procedure (3.4)–(3.6), the following can be derived.

(a) $M = 2$

(b) $M = 3$

(c) $M = 10$
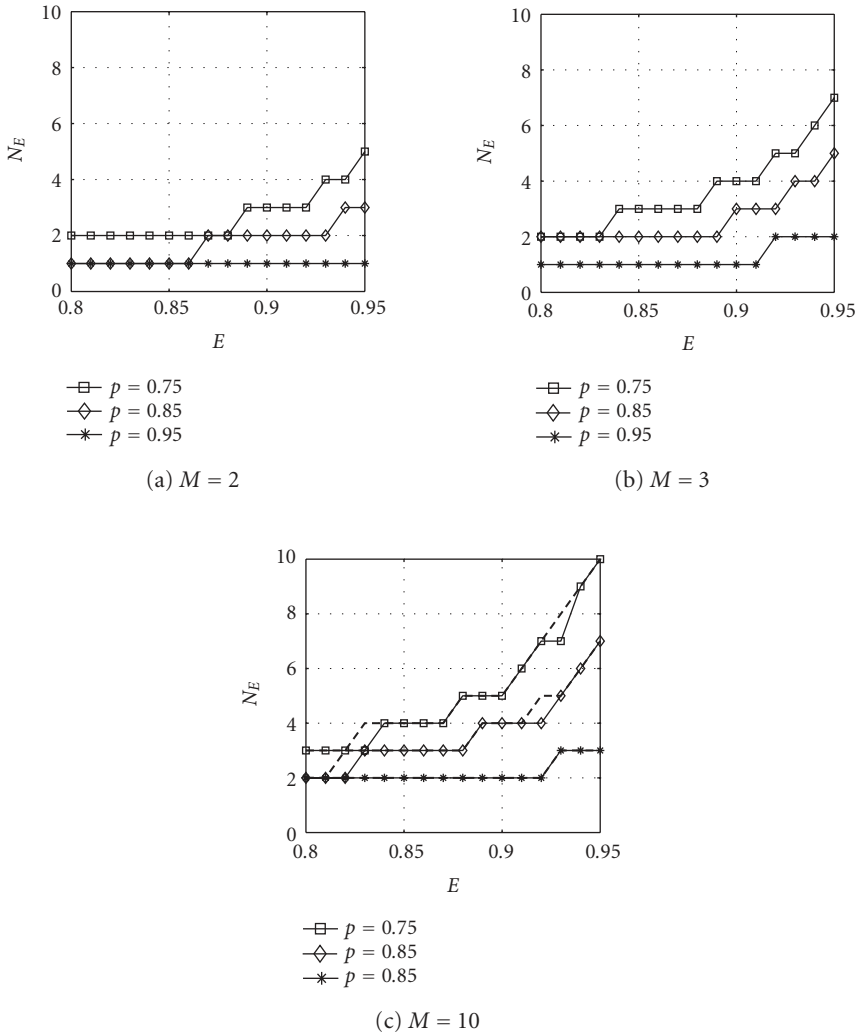
Figure 4.2. Lean buffering as a function of line efficiency.

PROPOSITION 4.2. *The lean buffering in serial lines defined by assumptions* (i)–(vi) *with* $M = 3$ *and* $p_1 = p_2 = p_3 = p$ *is given by*

$$N_E(M = 3) = \left\lceil \frac{\ln((1 - \sqrt{E})/(1 - E))}{\ln((1 - p)\sqrt{E}/(1 - p\sqrt{E}))} \right\rceil. \qquad (4.6)$$

For the proof, see the appendix.

The behavior of this N_E is illustrated in Figures 4.1(b) and 4.2(b). Obviously, for most values of p, the lean buffer capacity is increased, as compared with the case of $M = 2$,

and the range of p's, where JIT is possible, is decreased. For instance, if $p = 0.95$, JIT is acceptable for $E < 0.91$, while for $p = 0.85$ it is acceptable for $E < 0.77$.

4.3. M > 3-machine lines. For $M > 3$, no closed-form expression for the lean buffering can be derived. However, using aggregation procedure (3.4)–(3.6), the following estimate of N_E can be obtained.

PROPOSITION 4.3. *The lean buffering in serial lines defined by assumptions* (i)–(vi) *with* $M > 3$ *and* $p_i = p$, $i = 1, \ldots, M$, *is given by*

$$\hat{N}_E(M > 3) = \left\lceil \frac{\ln\left((1 - E - \hat{Q})/(1 - E)(1 - \hat{Q})\right)}{\ln\left((1 - p)(1 - \hat{Q})/(1 - p(1 - \hat{Q}))\right)} \right\rceil, \tag{4.7}$$

where $\hat{Q} = \hat{Q}(p, M, E)$ *is defined as*

$$\hat{Q} = 1 - E^{(1/2)[1 + (M - 3/M - 1)^{M/4}]} + \left(E^{(1/2)[1 + (M - 3/M - 1)^{M/4}]} - E^{(M - 2/M - 1)}\right) \exp\left\{-\frac{E^{1/(M-1)} - p}{(1 - E)^{(1/E)^{2E}}}\right\}. \tag{4.8}$$

For the proof, see the appendix.

The accuracy of estimates (4.7) and (4.8) has been evaluated numerically by calculating the exact value of N_E (using the aggregation procedure (3.4)–(3.6)) and comparing it with $\hat{N}_E$ as follows:

$$\Delta_E = \frac{\hat{N}_E - N_E}{N_E} \times 100\%. \tag{4.9}$$

The values of Δ_E have been calculated for $p \in [0.85, 0.9, 0.95]$, $M \in [5, 10, 15, 20, 25, 30]$, and $E \in [0.85, 0.9, 0.95]$. It turned out that $\Delta_E = 0$ for all combinations of these parameters except when $\{p = 0.85, M = 5, E = 0.85\}$, where it is equal to 50%. Thus, we conclude that $\hat{N}_E$ provides a sufficiently accurate estimate of N_E.

The behavior of $\hat{N}_E$ for $M = 10$ is illustrated in Figures 4.1(c) and 4.2(c). Clearly, the buffer capacity is increased as compared with $M = 3$, and JIT operation becomes unacceptable for all values of p and E analyzed.

Using (4.7) and (4.8), the behavior of lean buffering as a function of M can be analyzed. This is illustrated in Figure 4.3. Interestingly, and to a certain degree unexpectedly, $\hat{N}_E$ is constant for all $M > 10$. This implies that the lean buffering appropriate for lines with 10 machines is also appropriate for lines with any larger number of machines. Based on this observation, the following rule of thumb for selecting lean buffering can be formulated. In serial production lines defined by assumption (i)–(vi) with $M \geq 10$, the capacity of the lean buffering can be selected as shown in Table 4.1.

(a) $p = 0.85$

(b) $p = 0.90$

(c) $p = 0.95$

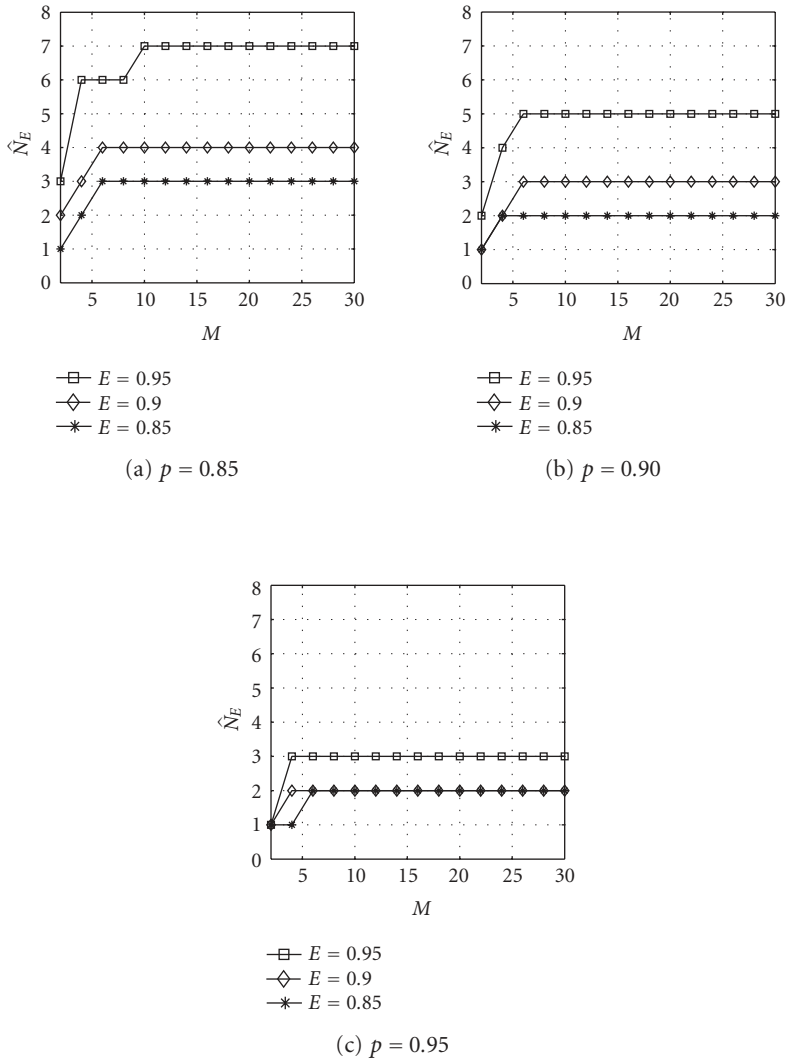Figure 4.3. Lean buffering $\hat{N}_E$ as a function of the number of machines in the system.

Table 4.1. Rule of thumb for selecting lean buffer capacity.

	$E = 0.85$	$E = 0.90$	$E = 0.95$
$p = 0.85$	3	4	7
$p = 0.90$	2	3	5
$p = 0.95$	2	2	3

5. Lean buffering in lines with nonidentical machines

5.1. Two-machine lines. In the case of nonidentical two-machine lines, as it follows from (2.1) and (3.1), the equation that defines N_E becomes

$$\text{PR} = \text{PR}_\infty E = p_2 [1 - Q(p_1, p_2, N_E)] = p_2 \left[1 - \frac{(1-p_1)(1-\alpha)}{1-(p_1/p_2)\alpha^{N_E}} \right], \qquad (5.1)$$

where

$$\text{PR}_\infty = \min (p_1, p_2), \qquad (5.2)$$

$$\alpha = \frac{p_1(1-p_2)}{p_2(1-p_1)}. \qquad (5.3)$$

From here we obtain the following.

PROPOSITION 5.1. *The lean buffering in serial lines defined by assumptions* (i)–(vi) *with* $M = 2$ *is given by*

$$N_E(p_1, p_2) = \left\lceil \frac{\ln \{(p_2/p_1)[(p_1 - E\text{PR}_\infty)/(p_2 - E\text{PR}_\infty)]\}}{\ln \alpha} \right\rceil, \qquad (5.4)$$

where PR_∞ *and* α *are defined by (5.2) and (5.3), respectively.*

Figure 5.1 illustrates the behavior of N_E as a function of p_1 for various values of p_2 and E, while Figure 5.2 shows N_E as a function of E for various p_1 and p_2. From these figures, we conclude the following.

(α) For p_2 sufficiently large, JIT operation is acceptable for all values of p_1 and E.

(β) For small p_2, JIT is acceptable only when p_1 is sufficiently large. For instance, if $p_2 = 0.75$, JIT represents LBC only if $p_1 > 0.88$ for $E = 0.9$ and $p_1 > 0.94$ for $E > 0.95$.

(γ) The maximum of N_E tends to take place when $p_1 = p_2$.

Intuitively, it is expected that the lean buffering in a line $\{p_1, p_2\}$ is the same as in the reversed line, that is, $\{p_2, p_1\}$. It turns out that this is indeed true as stated below.

PROPOSITION 5.2. *Lean buffer capacity has the property of reversibility, that is,*

$$N_E(p_1, p_2) = N_E(p_2, p_1). \qquad (5.5)$$

Proof. The proof follows immediately from (5.4) by observing that $\alpha(p_1, p_2)$ is equal to $1/\alpha(p_2, p_1)$ and $\text{PR}(p_1, p_2)$ is equal to $\text{PR}(p_2, p_1)$. □

5.2. M > 2-machine lines. Exact formulas for LBC in the case of $M > 2$ are all but impossible to derive. Therefore, we limit our attention to estimates of $N_{i,E}$. These estimates are obtained based on both closed formulas (4.5)–(4.8), (5.4), and recursive calculations. Each of them is described below.

(a) $E = 0.85$

(b) $E = 0.90$

(c) $E = 0.95$
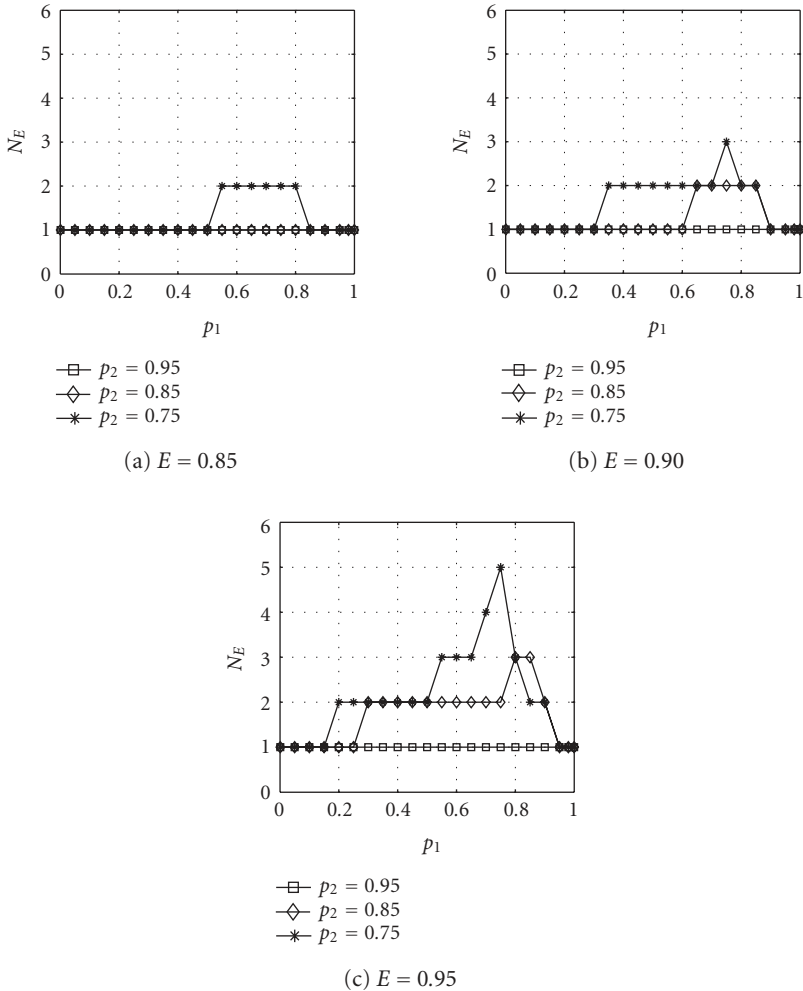
Figure 5.1. Lean buffering as a function of the first machine efficiency.

5.2.1. Closed-formula approaches. The following four methods have been investigated.

(I) *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 (5.4). This results in the sequence of buffer capacities denoted as

$$N_{1,E}^{\mathrm{I}},\dots,N_{M-1,E}^{\mathrm{I}}. \tag{5.6}$$

(II) *Global pairwise approach.* It is based on applying formula (5.4) 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}.
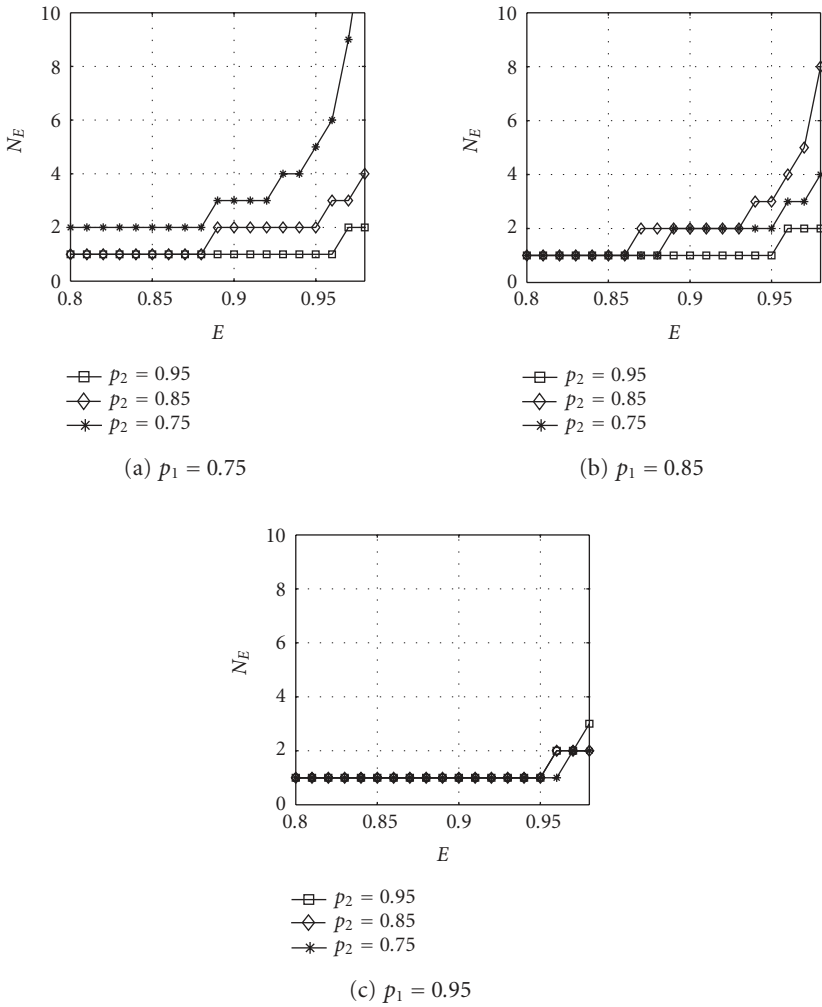
Figure 5.2. Lean buffering as a function of line efficiency.

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

$$\hat{p}_i := \min\{p_i, p_{i+1}\}, \quad i = 1,\ldots,M - 1, \tag{5.7}$$

and select LBC using formula (4.5) with $p = \hat{p}_i$. This results in the sequence of buffer capacities

$$N^{\mathrm{III}}_{1,E},\ldots,N^{\mathrm{III}}_{M-1,E}. \tag{5.8}$$

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

$$\hat{p} := \min\{p_1, p_2, \ldots, p_M\} \tag{5.9}$$

and select the buffer capacity, denoted as N_E^{IV}, using expressions (4.7) and (4.8). Due to the monotonicity of PR with respect to the machine efficiency and buffer capacity, this approach provides an upper bound of LBC:

$$N_{i,E} \le N_E^{IV}, \quad i = 1, \ldots, M - 1. \tag{5.10}$$

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 $\hat{p} = \min\{p_1, \ldots, p_M\}$, 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).

To analyze the performance of approaches (I)–(IV), we considered 100 000 lines formed by selecting M and p_i randomly and equiprobably from the sets

$$M \in \{4, 5, \ldots, 30\}, \tag{5.11}$$

$$0.70 \le p \le 0.97. \tag{5.12}$$

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

$$0.80 \le E \le 0.98. \tag{5.13}$$

For each kth line thus formed, we calculated the vector of buffer capacities,

$$\mathbf{N}_k^j = \begin{bmatrix} N_{1,k}^j \\ N_{2,k}^j \\ \ldots \\ N_{M-1,k}^j \end{bmatrix}, \quad k = 1, \ldots, 100\,000, \ j = (I), (II), (III), (IV), \tag{5.14}$$

using the four approaches introduced above. The subscripts of $N_{i,k}^j$ represent ith buffer, $i \in \{1, \ldots, M_{k-1}\}$, of the kth line, $k \in \{1, 2, \ldots, 100\,000\}$; the superscript $j \in \{(I), (II), (III), (IV)\}$ represents the approach used for this calculation. In addition, we calculated the production rate, PR_k^j, and the efficiency, E_k^j, using expressions (3.4)–(3.9) and (2.1), respectively.

The efficacy of approaches (I)–(IV) has been characterized by the following two metrics.

Table 5.1. Performance characteristics of approaches (I)–(IV).

Method	(I)	(II)	(III)	(IV)
N_{ave}^{j}	2.0	6.2	5.3	7.2
Δ^{j}	97.3	0.1	0.1	0.0

(1) The average buffer capacity per machine:

$$N_{\text{ave}}^{j} = \frac{1}{K} \sum_{k=1}^{K} N_k^j, \qquad (5.15)$$

where $K = 100\,000$ and

$$N_k^j = \frac{1}{M_k - 1} \sum_{i=1}^{M_k-1} N_{i,k}^j. \qquad (5.16)$$

(2) The frequency of E_k^j being less than the desired efficiency E_k:

$$\Delta^j = \frac{1}{K} \sum_{k=1}^{K} \text{Sg}(E_k - E_k^j) \cdot 100\%, \qquad (5.17)$$

where $K = 100\,000$ and

$$\text{Sg}(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0. \end{cases} \qquad (5.18)$$

The results are given in Table 5.1. Clearly, approach (I) leads to the smallest average buffer capacity but, unfortunately, almost always results in line efficiency less than desired. Thus, a "local" selection of LBC (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 2–3 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 1000 lines for each $M \in \{4, 6, 8, 10, 15, 20, 25, 30, 80\}$ by selecting p_i's and E's randomly and equiprobably from sets (5.12) and (5.13), respectively. For each of these lines, we calculated buffer capacities using approaches (I)–(IV) and evaluated the performance metrics (5.15) and (5.17). The results are shown in Table 5.2. Examining these data, we conclude the following.

(α) The local pairwise approach in most 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 N_{ave} and Δ. For $M \leq 10$, it outperforms approach (III) from the point of view of N_{avg}. However, it is quite sensitive to M: N_{ave} increases substantially with M.

Table 5.2. Effect of M on the performance of approaches (I)–(IV).

M	4	6	8	10	15	20	25	30	80
$N_{\text{ave}}^{\text{I}}$	1.7	1.9	2.0	2.0	2.0	2.1	2.0	2.1	2.1
$N_{\text{ave}}^{\text{II}}$	3.0	4.2	4.8	5.2	5.9	6.4	6.5	7.1	7.6
$N_{\text{ave}}^{\text{III}}$	4.2	5.0	5.2	5.2	5.2	5.4	5.3	5.7	5.6
$N_{\text{ave}}^{\text{IV}}$	5.0	6.3	6.7	6.8	7.0	7.4	7.3	7.9	7.9
Δ^{I}	88.7	92.8	95.6	97.1	98.4	98.0	99.2	99.3	100.0
Δ^{II}	3.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Δ^{III}	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.3
Δ^{IV}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

(γ) The local upper-bound approach is less sensitive to M and outperforms approach (II) for $M > 10$.

(δ) The global upper-bound approach substantially overestimates the LBC.

Based on the above, it is recommended to use the global pairwise approach in systems with $M \leq 10$ and local upper-bound approach in systems with $M > 10$.

5.2.2. 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 (3.4)–(3.9), calculate the production 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

$$N_{1,E}^{\text{V}},\dots,N_{M-1,E}^{\text{V}}. \tag{5.19}$$

(VI) *Bottleneck-based approach.* Consider a production line with the buffer capacity calculated according to approach (I) but rounded down in formula (5.4) 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 3, 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

$$N_{1,E}^{\text{VI}},\dots,N_{M-1,E}^{\text{VI}}. \tag{5.20}$$

Table 5.3. Performance characteristics of approaches (I)–(VI) in 5-machine lines.

Approaches	(I)	(II)	(III)	(IV)	(V)	(VI)
			(a) $E = 0.80$			
N^j_{ave}	1.5	2.2	2.4	2.9	1.4	1.5
Δ^j	27.8	0.2	0.0	0.0	0.0	0.0
t^j	8	8	8	8	59	28
			(b) $E = 0.85$			
N^j_{ave}	1.7	2.7	3.0	3.6	1.7	1.8
Δ^j	36.0	0.2	0.0	0.0	0.0	0.0
t^j	8	8	8	8	91	38
			(c) $E = 0.90$			
N^j_{ave}	2.2	3.6	4.2	5.0	2.0	2.3
Δ^j	33.8	0.0	0.0	0.0	0.0	0.0
t^j	6	6	6	6	91	27
			(d) $E = 0.95$			
N^j_{ave}	3.2	6.0	7.9	9.5	2.8	3.3
Δ^j	25.5	0.0	0.0	0.0	0.0	0.0
t^j	6	6	6	6	150	29

Clearly, approach (V) results 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 (V) with (VI), the computation time should be taken into account. This additional performance metric is defined as the total computer time necessary to carry out the computation

$$t^j = t^j_{end} - t^j_{start}, \tag{5.21}$$

where t^j_{start} and t^j_{end} are the times (in seconds) of the beginning and the end of the computation.

Based on performance metrics (5.15), (5.17), and (5.21), we compared approaches (I)–(VI) using the production systems generated by selecting p_i's randomly and equiprobably from set (5.12). The results are shown in Tables 5.3, 5.4, and 5.5. Specifically, Tables 5.3 and 5.4 present the results obtained using 5000 randomly generated 5- and 10-machine lines, respectively, while Table 5.5 is based on the analysis of 2000 randomly generated 15-machine lines. Examining these data, we conclude the following.

(α) Full-search approach, as expected, results in the smallest buffer capacity and the longest calculation time.

Table 5.4. Performance characteristics of approaches (I)–(VI) in 5-machine lines.

Approaches	(I)	(II)	(III)	(IV)	(V)	(VI)
(a) $E \in [0.80, 0.89]$						
N_{ave}^j	1.8	3.5	3.2	4.1	1.8	1.9
Δ^j	66.7	0.0	0.0	0.0	0.0	0.0
t^j	15	15	15	15	757	76
(b) $E \in [0.89, 0.98]$						
N_{ave}^j	3.2	8.0	8.2	10.6	2.7	3.2
Δ^j	48.5	0.0	0.0	0.0	0.0	0.0
t^j	26	26	26	26	2339	124

Table 5.5. Performance characteristics of approaches (I)–(VI) in 15-machine lines.

Approaches	(I)	(II)	(III)	(IV)	(V)	(VI)
(a) $E \in [0.80, 0.89]$						
N_{ave}^j	1.8	3.8	3.3	4.2	1.8	2.0
Δ^j	81.0	0.0	0.0	0.0	0.0	0.0
t^j	18	18	18	18	2452	107
(b) $E \in [0.89, 0.98]$						
N_{ave}^j	3.2	9.0	8.2	10.9	2.6	3.2
Δ^j	57.6	0.0	0.0	0.0	0.0	0.0
t^j	24	24	24	24	5837	138

(β) Approaches (II)–(IV), being based on closed-form expressions, are much faster than (V) (up to two orders of magnitude for long lines) but lead to the average buffering 2–3 times larger than that of (V).

(γ) Approach (VI) provides a good tradeoff between the calculation time and buffer capacity. In long lines, it is about 20 times faster than (V) and results in the average buffering almost the same as (V) (about 10% difference). Also, it is about 6 times slower than (I)–(IV) but gives buffering 2–3 times smaller than (II)–(IV).

5.2.3. 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 5.6 along with the desired line efficiency. The estimates of LBC for each of these lines, calculated using approaches (I)–(VI), are shown in Tables 5.7, 5.8, 5.9, and 5.10; these tables include also the resulting line efficiency, E^j. These examples clearly show the efficacy of the bottleneck approach, which is based on the closed formula (5.4) and the BN identification method described in Section 3.

Table 5.6. Desired line efficiencies and machine parameters of case studies.

Line	E	m_1	m_2	m_3	m_4	m_5
1	0.80	0.78	0.88	0.75	0.91	0.83
2	0.85	0.79	0.84	0.85	0.94	0.76
3	0.90	0.72	0.85	0.74	0.82	0.84
4	0.95	0.77	0.87	0.90	0.90	0.72

Table 5.7. LBC estimates for line 1.

Buffer	b_1^j	b_2^j	b_3^j	b_4^j	E^j
Desired	—	—	—	—	0.80
N_i^{I}	1	1	1	1	0.71
N_i^{II}	2	2	2	2	0.90
N_i^{III}	3	3	3	2	0.96
N_i^{IV}	3	3	3	3	0.96
N_i^{V}	1	2	2	1	0.83
N_i^{VI}	1	2	2	1	0.83

Table 5.8. LBC estimates for line 2.

Buffer	b_1^j	b_2^j	b_3^j	b_4^j	E^j
Desired	—	—	—	—	0.85
N_i^{I}	2	2	1	1	0.84
N_i^{II}	3	3	3	3	0.99
N_i^{III}	3	2	2	3	0.96
N_i^{IV}	3	3	3	3	0.99
N_i^{V}	1	2	2	1	0.85
N_i^{VI}	1	2	2	1	0.85

Table 5.9. LBC estimates for line 3.

Buffer	b_1^j	b_2^j	b_3^j	b_4^j	E^j
Desired	—	—	—	—	0.90
N_i^{I}	2	2	3	3	0.90
N_i^{II}	4	4	4	4	0.97
N_i^{III}	5	5	5	4	0.99
N_i^{IV}	5	5	5	5	0.99
N_i^{V}	2	3	2	2	0.90
N_i^{VI}	2	3	3	2	0.92

Table 5.10. LBC estimates for line 4.

Buffer	b_1^j	b_2^j	b_3^j	b_4^j	E^j
Desired	—	—	—	—	0.95
N_i^{I}	3	3	4	2	0.98
N_i^{II}	5	5	5	5	1.00
N_i^{III}	9	5	4	10	1.00
N_i^{IV}	10	10	10	10	1.00
N_i^{V}	2	2	2	2	0.96
N_i^{VI}	2	2	3	2	0.97

6. Conclusions

For serial production lines with *identical* Bernoulli machines, this paper provides closed-from expressions (4.7) and (4.8) for calculating the lean buffer capacity as a function of the line and machine efficiency and the number of machines in the system. For *nonidentical* machines, closed-form expression (5.4) has been derived only for two-machine lines. For longer lines, several approximations 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

 (i) if closed formulas are to be used for selecting lean buffering, the global pairwise and the local upper-bound approaches are recommended for $M \le 10$ and $M > 10$, respectively;

 (ii) if recursive approaches are to be employed, the bottleneck-based approach is preferred.

Appendix

We first prove Proposition 4.3 and then specialize it to Proposition 4.2.

Proof of Proposition 4.3. From (3.9) with $i = M - 1$, we obtain

$$\mathrm{PR} = p_M^b \left[1 - Q\left(p_{M-1}^f, p_M^b, N_{M-1} \right) \right] = p \left[1 - Q\left(p_{M-1}^f, p, N_E \right) \right]. \tag{A.1}$$

Using (2.1), (3.2), and (3.3), we find

$$1 - E = Q\left(p_{M-1}^f, p, N_E \right) = \frac{\left(1 - p_{M-1}^f \right)(1 - \alpha)}{1 - \left(p_{M-1}^f / p \right) \alpha^{N_E}}, \tag{A.2}$$

$$\alpha = \frac{p_{M-1}^f (1 - p)}{p \left(1 - p_{M-1}^f \right)}. \tag{A.3}$$

Solving (A.2) for N_E yields

$$N_E = \frac{\ln \left[\left(p / p_{M-1}^f \right) \left(\left(E - p_{M-1}^f - \alpha + \alpha p_{M-1}^f \right) / (E - 1) \right) \right]}{\ln \alpha}. \tag{A.4}$$

The steady state of (3.4) gives

$$p^f_{M-1} = p\left[1 - Q(p^f_{M-2}, p^b_{M-1}, N_E)\right] = p[1 - Q], \tag{A.5}$$

where

$$Q = Q(p^f_{M-2}, p^b_{M-1}, N). \tag{A.6}$$

Substituting (A.5) into (A.4) finally yields

$$N_E = \left\lceil \frac{\ln\left((1 - E - Q)/(1 - E)(1 - Q)\right)}{\ln\left((1 - p)(1 - Q)/(1 - p(1 - Q))\right)} \right\rceil. \tag{A.7}$$

Following the derivation of $\hat{Q}$ for exponential machines in [2], we approximate Q as

$$\hat{Q} = a + b\exp\left\{-\frac{\alpha - p}{\beta}\right\}, \tag{A.8}$$

where

$$a = 1 - E^{(1/2)[1+((M-3)/(M-1))^{M/4}]},$$

$$b = E^{(1/2)[1+((M-3)/(M-1))^{M/4}]} - E^{((M-2)/(M-1))}, \tag{A.9}$$

$$\alpha = E^{1/(M-1)}.$$

By trial and error, we select β as

$$\beta = (1 - E)^{(1/E)^{2E}}. \tag{A.10}$$

Thus, the approximation for Q is

$$\hat{Q} = 1 - E^{(1/2)[1+((M-3)/(M-1))^{M/4}]}$$

$$+ \left(E^{(1/2)[1+((M-3)/(M-1))^{M/4}]} - E^{((M-2)/(M-1))}\right)\exp\left\{-\frac{E^{1/(M-1)} - p}{(1 - E)^{(1/E)^{2E}}}\right\}. \tag{A.11}$$

Figure A.1 shows $\hat{Q}$ by broken lines and true Q, obtained through recursive aggregation, by solid lines. Tables A.1, A.2, and A.3 show that the accuracy of this approximation is within a 4% error, with the percent error as $(\hat{Q} - Q)/Q \times 100\%$. Thus, we can replace Q in (A.7) with $\hat{Q}$ to obtain (4.7). □

Proof of Proposition 4.2 In three-machine lines, (A.5) becomes

$$p^f_2 = p\left[1 - Q\left(p^f_1, p^b_2, N_E\right)\right] = p\left[1 - Q\left(p, p^b_2, N_E\right)\right], \tag{A.12}$$

and, in addition,

$$p^f_2 = p^b_2. \tag{A.13}$$

(a) $M = 10$, $E = 0.85$

(b) $M = 10$, $E = 0.90$

(c) $M = 10$, $E = 0.95$
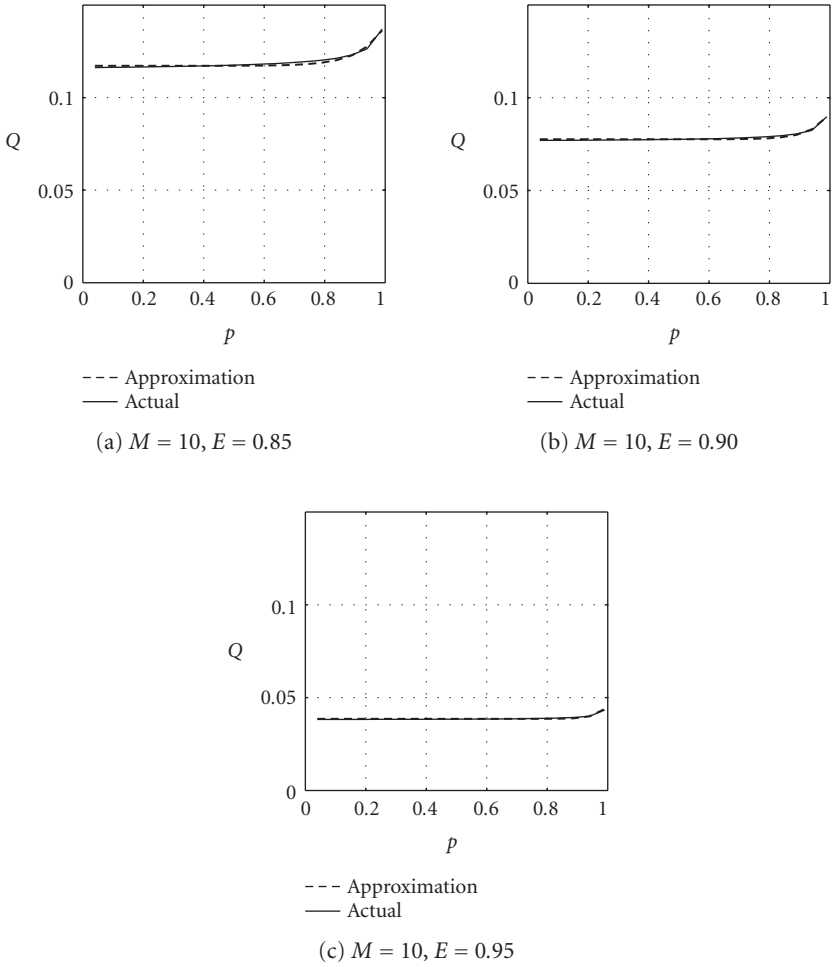
Figure A.1. Functions Q and $\hat{Q}$ versus p.

Therefore,

$$\mathrm{PR} = p_2^f\left[1 - Q\left(p, p_2^f, N_E\right)\right] = p\left[1 - Q\left(p, p_2^f, N\right)\right]^2 = p[1 - Q]^2, \qquad \text{(A.14)}$$

where Q was defined in (A.6).

From (A.14) and (2.1), we find

$$Q = 1 - \sqrt{E}. \qquad \text{(A.15)}$$

Substituting this into (A.5), and then substituting the resulting expression for p_{M-1}^f into (A.4), we obtain (4.6). □

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

p	$E = 0.85$	$E = 0.90$	$E = 0.95$
0.1	2.79	2.87	2.95
0.2	2.71	2.82	2.92
0.3	2.60	2.74	2.88
0.4	2.46	2.64	2.83
0.5	2.28	2.50	2.76
0.6	2.04	2.31	2.66
0.7	1.75	2.04	2.48
0.8	1.47	1.74	2.18
0.9	1.32	1.55	1.77

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

p	$E = 0.85$	$E = 0.90$	$E = 0.95$
0.1	0.69	0.75	0.82
0.2	0.52	0.63	0.76
0.3	0.30	0.48	0.68
0.4	0.05	0.29	0.57
0.5	−0.29	0.03	0.43
0.6	−0.68	−0.31	0.22
0.7	−1.06	−0.75	−0.10
0.8	−1.13	−1.11	−0.62
0.9	0.08	−0.41	−0.96

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

p	$E = 0.85$	$E = 0.90$	$E = 0.95$
0.1	2.63	2.74	3.72
0.2	2.40	2.59	3.18
0.3	2.14	2.40	2.68
0.4	1.80	2.15	2.55
0.5	1.36	1.83	2.37
0.6	0.82	1.37	2.11
0.7	0.22	0.76	1.69
0.8	−0.15	0.13	0.99
0.9	0.53	0.40	0.28

Acknowledgments

References

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

[2] E. Enginarlar, J. Li, and S. M. Meerkov, *How lean can lean buffers be?*, IIE Transactions **37** (2005), no. 4, 333–342.

[3] ⸺, *Lean buffering in serial production lines with non-exponential machines*, Stochastic Modeling of Manufacturing Systems (G. Liberopoulos, C. T. Papadopoulos, B. Tan, J. MacGregor Smith, and S. B. Gershwin, eds.), Springer, New York, 2006, pp. 29–53.

[4] E. Enginarlar, J. Li, S. M. Meerkov, and R. Q. Zhang, *Buffer capacity for accommodating machine downtime in serial production lines*, International Journal of Production Research **40** (2002), no. 3, 601–624.

[5] S. B. Gershwin and Y. Goldis, *Efficient algorithms for transfer line design*, Report LMP-95-005, Laboratory of Manufacturing and Productivity, MIT, Massachusetts, 1995.

[6] S. B. Gershwin and J. E. Schor, *Efficient algorithms for buffer space allocation*, Annals of Operations Research **93** (2000), no. 1–4, 117–144.

[7] D. Jacobs and S. M. Meerkov, *Mathematical theory of improvability for production systems*, Mathematical Problems in Engineering **1** (1995), no. 2, 95–137.

[8] C.-T. Kuo, J.-T. Lim, and S. M. Meerkov, *Bottlenecks in serial production lines: a system-theoretic approach*, Mathematical Problems in Engineering **2** (1996), no. 3, 233–276.

[9] J. MacGregor Smith and F. R. B. Cruz, *The buffer allocation problem for general finite buffer queueing networks*, IIE Transactions **37** (2005), no. 4, 343–365.

[10] J. O. McClain, R. Conway, W. Maxwell, and L. J. Thomas, *The role of work-in-process inventory in serial production lines*, Operations Research **36** (1988), no. 2, 229–241.

[11] H. Yamashita and T. Altiok, *Buffer capacity allocation for a desired throughput in production lines*, IIE Transactions **30** (1998), no. 10, 883–892.

A. B. Hu: Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122, USA
E-mail address: alexhu@umich.edu

S. M. Meerkov: Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122, USA
E-mail address: smm@eecs.umich.edu