Production Lead Time Problem: Formulation and Solution for Bernoulli Serial Lines *

Semyon M. Meerkov and Chao-Bo Yan

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109-2122

Abstract

This paper is intended to formulate the so-called production lead time problem as the problem of throughput optimization under a constraint on production lead time. A solution is provided for serial lines with Bernoulli machines and infinite buffers. Specifically, analytical formulas are derived for raw material release rates, which guarantee the desired lead time, while maximizing the throughput. Based on this solution, a method for selecting design parameters of kanban- and CONWIP-controlled systems, which ensure the desired lead time, is provided.

Keywords: Production systems, Throughput, Lead time, Raw material release, Kanban, CONWIP.

*This work was supported by NSF Grant No. CMMI-1160968.

1 Introduction

Production Systems Engineering ([1]) addresses two fundamental problems. The first one is the problem of throughput (*TP*) optimization. Its solution is provided by the improvability theory, i.e., *TP* optimization under the constraints on workforce and/or buffer capacity ([2]), and by bottleneck identification techniques, i.e., identification of a machine/operation that affects *TP* in the strongest manner ([3]). The second one is the problem of leanness, which is the problem of buffer capacity minimization under the constraint that *TP* takes a desired value. Its solution is provided by analytical formulas for the smallest buffer capacities, which are necessary and sufficient to ensure the desired *TP* ([4]).

In practice, there is another problem of critical importance – that of production lead time. The production lead time (also referred to as "flow time", "system cycle time", "process time", "residence time", etc.) is the average time a part spends in the system, being processed and waiting for processing. Excessively long lead time (*LT*) results in numerous quality, on-time delivery, and economic losses and, thus, should be avoided as much as possible. This problem is of particular importance in systems with "unlimited" buffers, where *LT* may also become unlimited.

The current paper is intended to formulate the production lead time problem as the problem of *TP* maximization under a constraint on *LT* and to provide its solution for serial lines with Bernoulli machines and infinite buffers. (Note that the Bernoulli reliability model implies that a machine produces a part during its cycle time with probability p and fails to do so with probability $1 - p$; as it is shown in [1], this model is applicable to operations, where machines downtime is relatively short and commensurate with the cycle time.)

The lead time in systems with infinite buffers can be controlled in two ways: (1) by limiting in-process inventory (e.g., using kanban or CONWIP system) or (2) by limiting release rates of raw materials. While the former have been analyzed in numerous publications (see, e.g., [5–18]), the latter remains practically unexplored. This paper is intended to contribute to the release rate approach and, in addition, provide a contribution to kanban and CONWIP systems. Specifically,

- we derive analytical formulas for *LT* as a function of the release rate and machine parameters;
- characterize achievable (i.e., feasible) sets of lead times;

- provide expressions for raw material release rates that guarantee any desired *LT* from the feasible set, while maximizing the throughput;

- based on the above results, offer a method for calculating the number of kanbans and the value of CONWIP that ensure a desired *LT* in kanban- and CONWIP-controlled systems.

Analytically, this work is based on a recursive aggregation procedure for performance evaluation of serial lines developed in [2]. In the current paper, this procedure is modified to account for infinite buffers. Since this recursive procedure provides estimates, rather than exact values of *TP* and work-in-process (*WIP*), the results obtained here are also approximate. The accuracy of these results is quantified by simulations and shown to be sufficiently high (well within 5%).

The outline of this paper is as follows: Section 2 formulates the model considered and the problems addressed. Sections 3 and 4 present the developments for Bernoulli serial lines with identical and non-identical machines, respectively. Deterministic (e.g., hourly) release is considered in Section 5, while the CONWIP and kanban systems are discussed in Section 6. Section 7 provides the conclusions and directions for future work. All proofs are given in the Appendix.

2 Modeling and Problem Formulation

Consider the serial line shown in Figure 2.1, where the circles represent the machines and the open rectangles are the buffers. While $m_1, m_2, \ldots, m_M$ and $b_1, b_2, \ldots, b_{M-1}$ are the usual producing machines and work-in-process buffers, respectively, m_0 and b_0 represent the mechanism of parts release and raw material storage (to indicate this, m_0 and b_0 are shown in gray). In other words, we model the raw material release rate by the efficiency, p_0, of the *release machine*, m_0. As mentioned above, the efficiency, p, of a Bernoulli machine is the probability to produce a part during a cycle time, while $1 - p$ is the probability of failing to do so. Thus, controlling $0 < p_0 < 1$, one can control the release rate in the system; in this case, the occupancy of buffer b_0 represents the raw material available for production.

To formalize this modeling approach, introduce the following assumptions:

(i) The system consists of M producing machines $m_1, m_2, \ldots, m_M$, a release machine m_0, $M - 1$

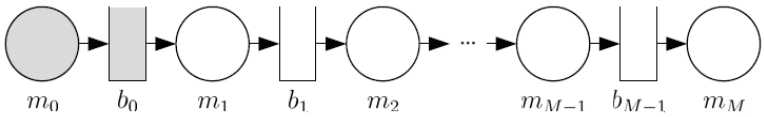

Figure 2.1: Serial production line with a release machine

work-in-process buffers $b_1, b_2, \ldots, b_{M-1}$, and the raw material buffer b_0. All machines have identical cycle time, τ, and obey the Bernoulli reliability model. All buffers are of infinite capacity.

(ii) While the efficiencies of the producing machines, i.e., $p_1, p_2, \ldots, p_M$, are fixed, the efficiency of the release machine, p_0, is free and can be selected at will.

(iii) The discrete event model is assumed, i.e., the time is slotted with the slot duration τ; the status of the machines (up or down) is determined at the beginning of each time slot, and the state (occupancy) of the buffers is determined at the end of each time slot.

(iv) A machine is starved, if the buffer in front of it is empty; a machine is blocked if the buffer after it is full and the subsequent machine does not take the material from this buffer; m_0 is never starved; m_M is never blocked.

(v) Machine failures are time-dependent, i.e., a machine can be down if it is blocked or starved.

Given the production system defined by these assumptions, the production lead time problem is formulated as follows: *determine the release rate of raw material, p_0^*, such that $LT = LT(p_0^*, p_1, \ldots, p_M)$ takes a desired value, while $TP = TP(p_0^*, p_1, \ldots, p_M)$ is maximized.*

Below we provide a solution of this problem. We begin with serial lines having identical producing machines and then extend the results to non-identical machines. For both identical and non-identical machines, we derive analytical expressions for the lead time as a function of machine parameters, evaluate the set of feasible lead times and, on this basis, derive analytical expressions for release rates that ensure the desired lead time and maximize the throughput.

3 Solution of Production Lead Time Problem for Serial Lines with Identical Machines

3.1 Lead time as a function of machine parameters

Theorem 3.1 *Consider a Bernoulli serial line defined by assumptions (i)-(v). Assume that all producing machines are identical, i.e.,* $p_i = p$, $i = 1, 2, \ldots, M$, *and the release machine is less efficient than the producing machines, i.e.,* $p_0 < p$. *Then, an estimate of the lead time (in units of cycle time,* τ*) is given by*

$$\widehat{LT} = \frac{M(1 - p_0)}{p - p_0}. \tag{3.1}$$

Proof: See the Appendix.

Thus, $\widehat{LT}$ is increasing linearly in M and tends to infinity hyperbolically as $p_0 \to p$. To further investigate the behavior of $\widehat{LT}$ as a function of the release rate and machine efficiency, introduce the following parameterizations:

$$\rho := \frac{p_0}{p}, \tag{3.2}$$

$$\widehat{lt} := \frac{\widehat{LT}}{M}. \tag{3.3}$$

Clearly, $0 < \rho < 1$ is the *relative workload* imposed on the system by the release, and $\widehat{lt} > 1$ is the *relative lead time*, i.e., the lead time in units of the smallest possible lead time in the system (i.e., M). Then, (3.1) can be re-written as follows:

$$\widehat{lt} = \frac{p^{-1} - \rho}{1 - \rho}. \tag{3.4}$$

The accuracy of this estimate is illustrated in Table 3.1, where $lt = \frac{LT}{M}$ is obtained by simulating 10-machine lines ($20,000$ time slots of warm-up time, $220,000$ time slots of simulation, and 20 repetitions). Clearly, the accuracy of (3.1) is quite high: up to the relative workload 0.99, it is well within 5%.

Figure 3.1 illustrates the behavior of $\widehat{lt}$ as a function of ρ for several p's. All curves in this

Table 3.1: Accuracy of estimate (3.4)

p	p_0	ρ	lt	$\widehat{lt}$	$\frac{\widehat{lt}-lt}{lt} \times 100\%$
0.5	0.2	0.40	2.66	2.67	0.38%
	0.4	0.80	5.99	6.00	0.17%
	0.45	0.90	10.90	11.00	0.92%
	0.495	0.99	97.79	101.00	3.28%
0.7	0.2	0.29	1.60	1.60	0
	0.4	0.57	2.00	2.00	0
	0.6	0.86	4.00	4.00	0
	0.65	0.92	7.02	7.00	−0.28%
	0.69	0.99	31.79	31.00	−2.49%
0.9	0.2	0.22	1.15	1.14	−0.87%
	0.4	0.44	1.20	1.20	0
	0.6	0.67	1.33	1.33	0
	0.8	0.89	2.00	2.00	0
	0.85	0.94	3.00	3.00	0
	0.89	0.99	10.58	11.00	3.97%

figure have a "knee" beyond which $\widehat{lt}$ grows extremely fast. It is of interest to characterize "safe" release rates, i.e., release rates below the knee. To accomplish this, consider the $(\rho, \widehat{lt})$-plane, where a unit interval of ρ-axis corresponds to $A > 1$ units of $\widehat{lt}$-axis (in Figure 3.1, $A = 20$). Introduce the *scaling ratio*, α, defined by

$$\alpha := \frac{1}{A} \tag{3.5}$$

and recall that the *curvature*, κ, of a twice differentiable function, $f(x)$, is given by (see [19])

$$\kappa(f(x)) = \frac{\left|f''_{xx}\right|}{(1 + f'^2_x)^{\frac{3}{2}}}. \tag{3.6}$$

Definition 3.1 *The knee, $\hat{\rho}^{knee}$, of $\widehat{lt}$ on the $(\rho, \widehat{lt})$-plane with the scaling ratio α is the point on $[0, 1)$, at which the curvature of $\alpha\widehat{lt}(\rho)$ reaches its maximum.*

Corollary 3.1 *Under the assumptions of Theorem 3.1,*

$$\hat{\rho}^{knee} = 1 - \sqrt{\alpha(p^{-1} - 1)}. \tag{3.7}$$

Proof: See the Appendix.

The pairs $(\hat{\rho}^{knee}, \widehat{lt}(\rho^{knee}))$ are indicated in Figure 3.1 by black dots. Thus, releasing raw material with the rate

$$p_0 < p(1 - \sqrt{\alpha(p^{-1} - 1)}) \tag{3.8}$$

results in $\widehat{lt}$ below the knee.

A precise characterization of the release rate that ensures the *desired* lead time is given next.

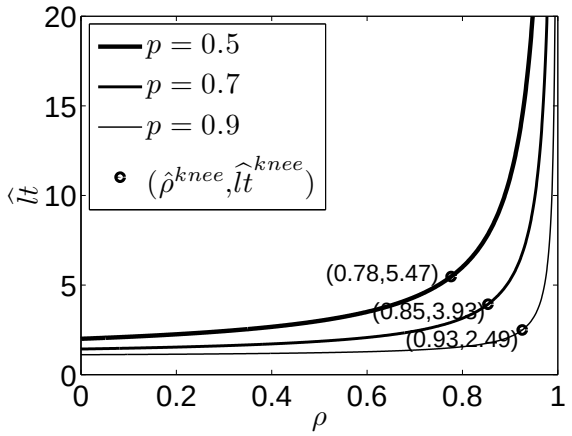

Figure 3.1: Relative lead time, $\widehat{lt}$, as a function of relative workload, ρ, and machine efficiency, p

3.2 Release rates as a function of desired lead time

From Theorem 3.1 we derive:

Corollary 3.2 *Under the assumptions of Theorem 3.1, the set of feasible lead times, $\mathscr{F}_{LT}$, is given by*

$$\widehat{LT} > Mp^{-1}. \tag{3.9}$$

Proof: See the Appendix.

This implies, in particular, that for low machine efficiency, lead time is substantially larger than the number of machines in the system. For instance, if $p = 0.5$, $\widehat{LT} > 2M$, no matter how low the release rate is.

Corollary 3.3 *Under the assumptions of Theorem 3.1, for any desired lead time, LT_d, satisfying (3.9), the release rate is given by*

$$\hat{p}_0^* = \frac{pLT_d - M}{LT_d - M}. \tag{3.10}$$

For this release rate,

$$\widehat{TP}^* = \hat{p}_0^*,\ \widehat{WIP}_i^* = \frac{\hat{p}_0^*(1-\hat{p}_0^*)}{p-\hat{p}_0^*},\ i = 0, 1, \ldots, M-1. \tag{3.11}$$

Proof: Follows immediately from (3.1), the proof of Theorem 3.1, and Little's law ([20]). ■

Note that since $p_0 > \hat{p}_0^*$ results in $\widehat{LT} > \widehat{LT}_d$, it follows from (3.11) that the release rate $\hat{p}_0^*$ maximizes the throughput under the lead constraint $LT = \widehat{LT}_d$ and, thus, solves the production lead time problem.

The behavior of $\hat{\rho}^*$ as a function of the desired relative lead time is illustrated in Figure 3.2 for several values of p. From this figure, we conclude that requiring small lt_d may lead to small $\hat{\rho}^*$, and therefore, low throughput. For instance, if $lt_d = 3$, the throughput is 0.25 for $p = 0.5$; 0.55 for $p = 0.7$; and 0.85 for $p = 0.9$; increasing lt_d to 10, leads to the throughput close to the machine efficiency; further increase of lt_d results in practically no throughput improvement. This leads to the following:

Rule-of-thumb 3.1 *In serial lines with identical Bernoulli machines, having lead time larger than* 10 *processing times, results in practically no throughput improvement.*

4 Solution of Production Lead Time Problem for Serial Lines with Non-identical Machines

4.1 Lead time as a function of machine parameters

Theorem 4.1 *Consider a Bernoulli serial line defined by assumptions (i)-(v). Assume that the release machine is less efficient than the producing machines, i.e., $p_0 < \min_{1 \leqslant i \leqslant M} p_i$. Then the estimate*

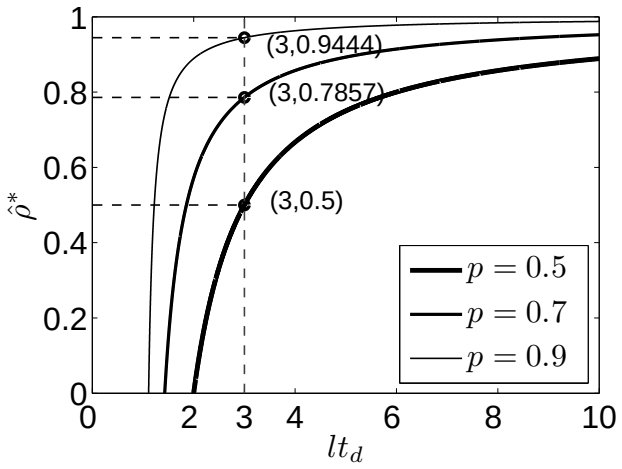

Figure 3.2: Relative workload, $\hat{\rho}^*$, as a function of the desired relative lead time, lt_d, and machine efficiency, p

of the lead time (in units of cycle time, τ) is given by

$$\widehat{LT} = \sum_{i=1}^{M} \frac{1 - p_0}{p_i - p_0}. \tag{4.1}$$

Proof: See the Appendix.

As one can see from this formula, $\widehat{LT}$ is monotonically increasing in M and tends hyperbolically to infinity as $p_0 \to \min\limits_{1 \leqslant i \leqslant M} p_i$. Note also that $\widehat{LT}$ is independent of the machine position in the line and, thus, the reversibility law ([1]) holds as far as $\widehat{LT}$ is concerned.

To investigate the accuracy of estimate (4.1), introduce the notion of maximum relative workload,

$$\rho_{max} = \frac{p_0}{\min\limits_{1 \leqslant i \leqslant M} p_i}, \tag{4.2}$$

and consider three lines,

$$\begin{aligned} L_1 &: p = [0.89, 0.78, 0.84, 0.82, 0.88, 0.86, 0.82, 0.75, 0.87, 0.82], \\ L_2 &: p = [0.84, 0.87, 0.89, 0.86, 0.78, 0.81, 0.89, 0.89, 0.81, 0.88], \\ L_3 &: p = [0.76, 0.80, 0.87, 0.75, 0.77, 0.78, 0.78, 0.84, 0.79, 0.78], \end{aligned} \tag{4.3}$$

where the machine efficiency, p_i, is selected randomly and equiprobably from the set $[0.75, 0.9]$. Using these lines, the accuracy of $\widehat{LT}$, evaluated based on the same simulation procedure as in

Section 3, is characterized in Table 4.1. Again, up to $\rho_{max} = 0.99$, the accuracy remains high.

Table 4.1: Accuracy of estimate (4.1)

p_0	ρ_{max}	line L_1			line L_2			line L_3		
		LT	$\widehat{LT}$	$\frac{\widehat{LT}-LT}{LT} \times 100\%$	LT	$\widehat{LT}$	$\frac{\widehat{LT}-LT}{LT} \times 100\%$	LT	$\widehat{LT}$	$\frac{\widehat{LT}-LT}{LT} \times 100\%$
0.1	0.13	12.39	12.32	−0.56%	11.98	12.00	0.17%	13.01	13.04	0.23%
0.2	0.27	12.74	12.70	−0.31%	12.26	12.31	0.41%	13.65	13.56	−0.66%
0.3	0.40	13.20	13.22	0.15%	12.76	12.74	−0.16%	14.37	14.29	−0.56%
0.4	0.53	13.98	14.00	0.14%	13.36	13.37	0.07%	15.43	15.42	−0.06%
0.5	0.67	15.26	15.28	0.13%	14.40	14.38	−0.14%	17.32	17.34	0.12%
0.6	0.80	17.68	17.84	0.90%	16.30	16.28	−0.12%	21.43	21.44	0.05%
0.7	0.93	26.39	26.28	−0.42%	21.38	21.39	0.05%	36.78	36.78	0
0.74	0.99	52.75	52.61	−0.27%	27.82	27.75	−0.25%	81.54	81.30	−0.29%

4.2 Release rates as a function of desired lead time

From Theorem 4.1 we derive:

Corollary 4.1 *Under the assumptions of Theorem 4.1, the set of feasible lead times, $\mathscr{F}_{LT}$, is given by*

$$\widehat{LT} > \sum_{i=1}^{M} p_i^{-1}. \tag{4.4}$$

Proof: See the Appendix.

Corollary 4.2 *Under the assumptions of Theorem 4.1, for any desired lead time, LT_d, satisfying (4.4), the release rate $\hat{p}_0^*$ that ensures this lead time is the unique real root less than $\min\limits_{1\leqslant i\leqslant M} p_i$ of the following M-th order polynomial equation:*

$$(LT_d - M)\prod_{i=1}^{M}(p_i - p_0) - \sum_{i=1}^{M}\Big((1-p_i)\prod_{j=1, j\neq i}^{M}(p_j - p_0)\Big) = 0, \tag{4.5}$$

and

$$\widehat{TP}^* = \hat{p}_0^*,\ \widehat{WIP}_i^* = \frac{\hat{p}_0^*(1-\hat{p}_0^*)}{p_{i+1} - \hat{p}_0^*},\ i = 0, 1, \ldots, M-1. \tag{4.6}$$

Proof: See the Appendix.

For the case of $M = 2$, equation (4.5) takes the form

$$(LT_d - 2)p_0^2 - [(LT_d - 1)(p_1 + p_2) - 2]p_0 + LT_d p_1 p_2 - (p_1 + p_2) = 0, \tag{4.7}$$

and, thus, the release rate is given by

$$\hat{p}_0^* = \frac{(LT_d - 1)(p_1 + p_2) - 2 - \sqrt{(LT_d - 1)^2(p_1 - p_2)^2 + 4(1 - p_1)(1 - p_2)}}{2(LT_d - 2)}. \tag{4.8}$$

4.3 Lower bound approach

Expression (4.5) and the subsequent evaluation of its roots may be too involved for practical applications. Therefore, we propose the following lower bound approach that can be useful in these situations:

Given a Bernoulli line with machines defined by $p_1, p_2, \ldots, p_M$, introduce

$$p_{min} := \min_{1 \leqslant i \leqslant M} p_i, \tag{4.9}$$

and, along with the original line, consider an auxiliary one with identical machines defined by p_{min}. Using Corollaries 3.3 and 4.2, select $\widehat{LT}_d$ in the feasible set for both the original and auxiliary lines and calculate the corresponding release rates ($\hat{p}_0^*$ for the original line and $\hat{p}_{0,min}^*$ for the auxiliary one). Clearly, due to the monotonicity of $\widehat{LT}$ with respect to machine efficiency,

$$\hat{p}_{0,min}^* \leqslant \hat{p}_0^*, \tag{4.10}$$

and, therefore,

$$\widehat{LT}(p_1, p_2, \ldots, p_M) \leqslant \widehat{LT}(p_{min}) = \widehat{LT}_d. \tag{4.11}$$

Thus, releasing raw material in the original line according to the lower bound, $\hat{p}_{0,min}^*$, does not result in a lead time longer than desired. Unfortunately, however, it may result in a lower throughput, since

$$\widehat{TP}(\hat{p}_{0,min}^*) \leqslant \widehat{TP}(\hat{p}_0^*). \tag{4.12}$$

To evaluate how much of throughput is lost due to raw material releasing according to the lower bound, consider the three lines defined in (4.3) and calculate $\hat{p}_0^*$ and $\hat{p}_{0,min}^*$ for various $\widehat{lt}$'s. The results are shown in Table 4.2. Clearly, throughput losses are large for small lt_d's (almost 50%

for $lt_d = 1.5$), but relatively insignificant for large lt_d (about 5% for $lt_d = 6$).

Table 4.2: Performance degradation due to lower bound release rates

lt_d	line L_1			line L_2			line L_3		
	$\hat{p}_0^*$	$\hat{p}_{0,min}^*$	$\frac{\hat{p}_0^* - \hat{p}_{0,min}^*}{\hat{p}_0^*} \times 100\%$	$\hat{p}_0^*$	$\hat{p}_{0,min}^*$	$\frac{\hat{p}_0^* - \hat{p}_{0,min}^*}{\hat{p}_0^*} \times 100\%$	$\hat{p}_0^*$	$\hat{p}_{0,min}^*$	$\frac{\hat{p}_0^* - \hat{p}_{0,min}^*}{\hat{p}_0^*} \times 100\%$
1.5	0.483	0.250	48.24%	0.541	0.340	37.15%	0.368	0.250	32.07%
2	0.643	0.500	22.24%	0.683	0.560	18.01%	0.574	0.500	12.89%
3	0.715	0.625	12.59%	0.747	0.670	10.31%	0.675	0.625	7.41%
4	0.732	0.667	8.88%	0.764	0.707	7.46%	0.708	0.667	5.79%
5	0.739	0.688	6.90%	0.770	0.725	5.84%	0.723	0.688	4.84%
6	0.742	0.700	5.66%	0.773	0.736	4.79%	0.731	0.700	4.24%

5 Deterministic Release

In some cases, random raw material release may be inconvenient in practical implementations. In such situations, results of Sections 3 and 4 can be used to define strategies for deterministic, e.g., hourly, release.

To model the hourly release, let the desired lead time be defined in minutes and denoted as $\mathscr{LT}_d$. Then the desired lead time in units of cycle time is given by

$$LT_d = \frac{\mathscr{LT}_d}{\tau}, \tag{5.1}$$

where τ is also in minutes. For example, if $\mathscr{LT}_d = 120$ min, $LT_d = 120$ if $\tau = 1$ min and $LT_d = 1200$ if $\tau = 0.1$ min. Given LT_d defined by (5.1), the corresponding release rate per cycle, $\hat{p}_0^*$, can be calculated using either Corollary 3.3 or 4.2. Then, the hourly release, $\mathscr{R}^*$, is defined as

$$\mathscr{R}^* = \lfloor H\hat{p}_0^*(LT_d) \rfloor, \tag{5.2}$$

where $\lfloor x \rfloor$ is the largest integer not greater than x and H is the number of cycles in an hour, i.e.,

$$H = \frac{60}{\tau}. \tag{5.3}$$

Releasing each hour the amount of raw material defined by (5.2), leads to the following in-

equalities:

$$\widehat{LT}(\hat{p}_0^*) < \widehat{LT}(\mathscr{R}^*) < \widehat{LT}(\hat{p}_0^*) + H, \tag{5.4}$$

where $\widehat{LT}(\mathscr{R}^*)$ and $\widehat{LT}(\hat{p}_0^*)$ are lead time estimates under hourly and per-cycle release, respectively. Multiplying these inequalities by τ gives

$$\widehat{\mathscr{LT}}_d < \widehat{\mathscr{LT}}(\mathscr{R}^*) < \widehat{\mathscr{LT}}_d + 60. \tag{5.5}$$

This indicates that, although releasing on the hourly basis results in lead times larger than desired, the difference becomes insignificant when $\mathscr{LT}_d \gg 60$ min.

6 Selecting Design Parameters of CONWIP and Kanban Systems to Ensure Desired Lead Time

Consider a CONWIP-controlled serial line, i.e., the system where a new part is released only when an old part has been processed by the last machine ([13]). Clearly, such a system maintains a constant number of parts, K^*. How should K^* be selected so that the throughput is maximized, while the lead time takes a desired value? To answer this question using the results derived above, assume that the desired lead time (in minutes) is $\mathscr{LT}_d$. Then, using (5.1) calculate the lead time in units of cycle time, LT_d. Based on this LT_d, calculate the release rate $\hat{p}_0^*$ using either Corollary 3.3 or 4.2. Under this release rate, the total *WIP* in the system can be evaluated using either (3.11) or (4.6). In other words,

$$\widehat{WIP}^* = M\hat{p}_0^* \frac{1-\hat{p}_0^*}{p-\hat{p}_0^*} \text{ for identical machine case,} \tag{6.1}$$

$$\widehat{WIP}^* = \hat{p}_0^* \sum_{i=1}^{M} \frac{1-\hat{p}_0^*}{p_i-\hat{p}_0^*} \text{ for non-identical machine case.} \tag{6.2}$$

Since under this *WIP* the throughput is maximized and the lead time takes the desired value

$\mathscr{LT}_d$, the number K^* in a CONWIP-controlled system should be selected as

$$K^* = \left\lfloor \widehat{WIP}^* \right\rfloor. \tag{6.3}$$

Assuming that raw material is always available in a kanban-controlled system, similar arguments lead to the conclusion that the number of kanbans that maximizes TP under a constraint on $\mathscr{LT}$ also should be selected as in (6.3).

The following examples illustrate these conclusions.

Consider the kanban- or CONWIP-controlled serial line with 10 identical Bernoulli machines having $p = 0.7$ and $\tau = 1$ min. Assume $\mathscr{LT}_d = 60$ min and, using (3.10), (6.1), and (6.3) calculate K^*, which turns out to be 38. Simulating the resulting system (for $50,000$ time slots with $25,000$ time slots of warm-up period and 20 repetitions) results in $\mathscr{LT}_{kanban} = \mathscr{LT}_{CONWIP} = 58.82$ min.

A similar result is obtained by simulating L_1 of (4.3) for the same $\mathscr{LT}_d$ and τ as above. In this case, K^* turns out to be 44 and the simulations result in $\mathscr{LT}_{kanban} = \mathscr{LT}_{CONWIP} = 58.64$ min.

Thus, selecting K^* using (6.3) indeed leads to an acceptable lead time performance of kanban- and CONWIP-controlled systems.

7 Conclusions and Future Research

This paper posed the problem of production lead time control as the problem of throughput optimization under the constraint that the lead time takes a desired value. The approach is based on limiting release rate of raw material so that the desired performance is achieved. This approach is of particular importance for small and mid-size manufacturing organizations, where neither hardware-constrained buffers nor kanban or CONWIP systems are available to limit the in-process inventory, but hourly release can be implemented relatively easily. For the case of kanban and CONWIP systems, however, this paper provided a method for calculating the number of kanbans and the value of CONWIP that ensure a desired LT.

Since this paper addressed only the case of serial lines with Bernoulli machines, future research on the lead time problem is abound. It includes

- extending the results to exponential and non-exponen- tial machines;
- extending the results to assembly systems;
- developing a solution of the production lead time problem for re-entrant lines; due to their complex dynamics (see [21]), it seems that the approach pursued in the current paper may not be appropriate, and a fundamentally new modeling approach would be necessary.

Appendix

The analysis of $\widehat{LT}$ reported in this paper is based on the recursive aggregation procedure described in [1]. For serial lines with $M+1$ Bernoulli machines defined by $p_0, p_1, \ldots, p_M$ and M buffers with capacity $N_0, N_1, \ldots, N_{M-1}$, the steady state of this procedure, p_i^f, $i = 1, 2, \ldots, M$, and p_i^b, $i = 0, 1, \ldots, M-1$, is the unique solution of the following system of transcendental equations:

$$\begin{aligned} p_i^f &= p_i[1 - Q(p_{i-1}^f, p_i^b, N_{i-1})],\ 1 \leqslant i \leqslant M, \\ p_i^b &= p_i[1 - Q(p_{i+1}^b, p_i^f, N_i)],\ 0 \leqslant i \leqslant M-1, \end{aligned} \tag{A.1}$$

with the boundary conditions $p_0^f = p_0$ and $p_M^b = p_M$ and

$$Q(x, y, N) = \begin{cases} \frac{1-x}{N+1-x}, & \text{if } x = y, \\ \frac{(1-x)(1-\alpha)}{1-\frac{x}{y}\alpha^N}, & \text{if } x \neq y, \end{cases} \tag{A.2}$$

$$\alpha = \frac{x(1-y)}{y(1-x)}. \tag{A.3}$$

The proofs of Theorems 3.1 and 4.1 are based on (A.1)-(A.3). Therefore, below we evaluate (A.2) and the solution of (A.1) for $N_i = \infty$ (Lemmas A.1 and A.2, respectively) and then prove the above mentioned theorems.

Lemma A.1 *Function $Q(x, y, N)$, defined by (A.2), (A.3), has the following limit:*

$$\lim_{N\to\infty} Q(x, y, N) = \begin{cases} 0, & \text{if } x \geqslant y, \\ 1 - \frac{x}{y}, & \text{if } x < y. \end{cases} \tag{A.4}$$

Proof: From (A.2),

- if $x = y$, it is clear that

$$\lim_{N\to\infty} Q(x, y, N) = 0; \tag{A.5}$$

- if $x > y$, then $\alpha > 1$, and, therefore,

$$\lim_{N\to\infty} Q(x, y, N) = 0; \tag{A.6}$$

- if $x < y$, then $\alpha < 1$, and, therefore,

$$\lim_{N\to\infty} Q(x, y, N) = (1 - x)(1 - \alpha) = 1 - \frac{x}{y}. \tag{A.7}$$

■

Lemma A.2 *Let $p_j = \min\limits_{1\leqslant i\leqslant M} p_i$. Then, for $N_i = \infty$, $i = 0, 1, \ldots, M - 1$, the unique solution of (A.1) is given by*

$$\begin{aligned} p_i^f &= \begin{cases} p_i, & \text{if } i < j, \\ p_j, & \text{if } i \geqslant j, \end{cases} \\ p_i^b &= \begin{cases} p_j, & \text{if } i \leqslant j, \\ p_i, & \text{if } i > j. \end{cases} \end{aligned} \tag{A.8}$$

Proof: First we show that (A.8) is the solution of (A.1) and then comment its uniqueness.

- If $i < j$, then based on (A.4) and (A.8), for the left- and right-hand side of the first equation of (A.1), we have, respectively,

$$p_i^f = p_i \tag{A.9}$$

and

$$p_i[1 - Q(p_{i-1}^f, p_i^b, \infty)] = p_i[1 - Q(p_i, p_j, \infty)] = p_i, \tag{A.10}$$

implying that (A.8) solve the first equation of (A.1) for $i < j$. Similarly, for the left- and right-hand side of the second equation of (A.1), we have

$$p_i^b = p_j \tag{A.11}$$

and

$$p_i[1 - Q(p_{i+1}^b, p_i^f, \infty)] = p_i[1 - Q(p_j, p_i, \infty)] = p_j, \tag{A.12}$$

implying that the second equation of (A.1) is also solved for $i < j$.

- If $i = j$, the left- and right-hand side of equations of (A.1) are

$$\begin{aligned} &p_i^f = p_j, \\ &p_i[1 - Q(p_{i-1}^f, p_i^b, \infty)] = p_j[1 - Q(p_{j-1}, p_j, \infty)] = p_j, \\ &p_i^b = p_j, \\ &p_i[1 - Q(p_{i+1}^b, p_i^f, \infty)] = p_j[1 - Q(p_{j+1}, p_j, \infty)] = p_j, \end{aligned} \tag{A.13}$$

  implying that (A.1) is solved for $i = j$.

- If $i > j$, the left- and right-hand side of equations of (A.1) are

$$\begin{aligned} &p_i^f = p_j, \\ &p_i[1 - Q(p_{i-1}^f, p_i^b, \infty)] = p_i[1 - Q(p_j, p_i, \infty)] = p_j, \\ &p_i^b = p_i, \\ &p_i[1 - Q(p_{i+1}^b, p_i^f, \infty)] = p_i[1 - Q(p_{i+1}, p_j, \infty)] = p_i, \end{aligned} \tag{A.14}$$

  which also implies that (A.1) is solved.

As far as the uniqueness of (A.8) is concerned, it follows directly from Theorem 4.2 of [1]. ■

Proof of Theorem 3.1: For the production line defined by assumptions (i)-(v) with $p_i = p$, $i = 1, 2, \ldots, M$ and $p_0 < p$, based on Lemma A.2 we obtain:

$$p_i^f = p_0,\ i = 0, 1, \ldots, M-1, \tag{A.15}$$

$$p_i^b = p,\ i = 1, 2, \ldots, M, \tag{A.16}$$

which, using Theorem 4.1 of [1], implies that the occupancy of each buffer is

$$\widehat{WIP}_0 = \frac{p_0(1-p_0)}{p-p_0} \tag{A.17}$$

and

$$\widehat{WIP}_i = \frac{p_i^f(1-p_i^f)}{p_{i+1}^b - p_i^f} = \frac{p_0(1-p_0)}{p-p_0},\ i = 1, 2, \ldots, M-1. \tag{A.18}$$

Thus, taking into account that

$$\widehat{PR} = p_0 \tag{A.19}$$

and using Little's law, from (A.17), (A.18) we obtain (3.1). ■

Proof of Corollary 3.1: Let

$$f(\rho) := \alpha\widehat{lt}(\rho) = \alpha\Big(\frac{p^{-1}-1}{1-\rho} + 1\Big). \tag{A.20}$$

Then

$$f'(\rho) = \frac{\alpha(p^{-1}-1)}{(1-\rho)^2},$$
$$f''(\rho) = \frac{2\alpha(p^{-1}-1)}{(1-\rho)^3},$$

and, therefore,

$$\kappa(f(\rho)) = \frac{\left|f''_{\rho\rho}\right|}{(1+f'^2_\rho)^{\frac{3}{2}}} = \frac{2\alpha(p^{-1}-1)}{\left[(1-\rho)^2 + \frac{\alpha^2(p^{-1}-1)^2}{(1-\rho)^2}\right]^{\frac{3}{2}}}. \tag{A.21}$$

Since

$$\rho^{knee} = \arg\max \kappa(f(\rho)), \tag{A.22}$$

from (A.21) we obtain:

$$\hat{\rho}^{knee} = 1 - \sqrt{\alpha(p^{-1} - 1)}. \tag{A.23}$$

■

Proof of Corollary 3.2: From (3.4) it follows that

$$\rho = 1 - \frac{p^{-1} - 1}{\widehat{lt} - 1}. \tag{A.24}$$

Since $0 < \rho < 1$, this implies that

$$\widehat{lt} > p^{-1}. \tag{A.25}$$

■

Proof of Corollary 3.2: From (3.4) it follows that

$$\rho = 1 - \frac{p^{-1} - 1}{\widehat{lt} - 1}. \tag{A.26}$$

Since $0 < \rho < 1$, this implies that

$$\widehat{lt} > p^{-1}. \tag{A.27}$$

■

Proof of Theorem 4.1: Similar to the proof of Theorem 3.1, with the only difference that, instead of equation (A.16), we have

$$p_i^b = p_i,\ i = 1, 2, \ldots, M \tag{A.28}$$

and, therefore,

$$\widehat{WIP}_0 = \frac{p_0(1 - p_0)}{p_1 - p_0}, \tag{A.29}$$

$$\widehat{WIP}_i = \frac{p_i^f(1 - p_i^f)}{p_{i+1}^b - p_i^f} = \frac{p_0(1 - p_0)}{p_{i+1} - p_0},\ i = 1, 2, \ldots, M - 1. \tag{A.30}$$

■

Proof of Corollary 4.1: Re-writing (4.1) as

$$\widehat{LT} - M = \sum_{i=1}^{M} \frac{1 - p_i}{p_i - p_0}, \tag{A.31}$$

and taking into account that $p_0 < \min\limits_{1 \leqslant i \leqslant M} p_i$, we observe that the right-hand side of (A.31) is a monotonically increasing function of p_0. Thus,

$$\widehat{LT} - M > \sum_{i=1}^{M} \frac{1 - p_i}{p_i}, \tag{A.32}$$

i.e., (4.4) holds. ■

Proof of Corollary 4.2: Under the assumptions of Theorem 4.1, for any desired lead time LT_d satisfying (4.4), the release rate $\hat{p}_0^*$ that ensures this lead time is a real root less than $\min\limits_{1 \leqslant i \leqslant M} p_i$ of the equation

$$LT_d = \sum_{i=1}^{M} \frac{1 - p_0}{p_i - p_0} \tag{A.33}$$

or

$$LT_d - M = \sum_{i=1}^{M} \frac{1 - p_i}{p_i - p_0}. \tag{A.34}$$

Since the right-hand side of (A.34) is monotonically increasing with p_0 when $0 < p_0 < \min\limits_{1 \leqslant i \leqslant M} p_i$, equation (A.34) has a unique real solution less than $\min\limits_{1 \leqslant i \leqslant M} p_i$ ensuring the desired lead time LT_d. Multiplying (A.34) by $\prod\limits_{j=1}^{M} (p_j - p_0)$, we have

$$(LT_d - M) \prod_{i=1}^{M} (p_i - p_0) = \sum_{i=1}^{M} \Big((1 - p_i) \prod_{j=1, j \neq i}^{M} (p_j - p_0) \Big), \tag{A.35}$$

i.e.,

$$(LT_d - M) \prod_{i=1}^{M} (p_i - p_0) - \sum_{i=1}^{M} \Big((1 - p_i) \prod_{j=1, j \neq i}^{M} (p_j - p_0) \Big) = 0. \tag{A.36}$$

In other words, for any desired lead time LT_d satisfying (4.4), the release rate $\hat{p}_0^*$ that ensures this lead time is the unique real root less than $\min\limits_{1 \leqslant i \leqslant M} p_i$ of the M-th order polynomial equation (4.5).

The statements on $\widehat{PR}^*$ and $\widehat{WIP}^*$ follow from the proof of Theorem 4.1. ■

References

[1] J. Li and S. M. Meerkov, *Production Systems Engineering*. New York: Springer, 2009.

[2] D. Jacobs and S. M. Meerkov, "A system-theoretic property of serial production lines: Improvability," *International Journal of Systems Science*, vol. 26, no. 4, pp. 755–785, 1995.

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

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

[5] Y. Sugimori, K. Kusunoki, F. Cho, and S. Uchikawa, "Toyota production system and kanban system materialization of just-in-time and respect-for-human system," *The International Journal of Production Research*, vol. 15, no. 6, pp. 553–564, 1977.

[6] J. L. Deleersnyder, T. J. Hodgson, H. Muller-Malek, and P. J. O'Grady, "Kanban controlled pull systems: An analytic approach," *Management Science*, vol. 35, no. 9, pp. 1079–1091, 1989.

[7] D. Mitra and I. Mitrani, "Analysis of a kanban discipline for cell coordination in production lines. I," *Management Science*, vol. 36, no. 12, pp. 1548–1566, 1990.

[8] B. J. Berkley, "A review of the kanban production control research literature," *Production and Operations Management*, vol. 1, no. 4, pp. 393–411, 1992.

[9] S. R. Tayur, "Structural properties and a heuristic for kanban-controlled serial lines," *Management Science*, vol. 39, no. 11, pp. 1347–1368, 1993.

[10] M. Di Mascolo, Y. Frein, and Y. Dallery, "An analytical method for performance evaluation of kanban controlled production systems," *Operations Research*, vol. 44, no. 1, pp. 50–64, 1996.

[11] M. S. Akturk and F. Erhun, "An overview of design and operational issues of kanban systems," *International Journal of Production Research*, vol. 37, no. 17, pp. 3859–3881, 1999.

[12] M. Lage, Jr. and M. G. Filho, "Variations of the kanban system: Literature review and classification," *International Journal of Production Economics*, vol. 125, no. 1, pp. 13–21, 2010.

[13] M. L. Spearman, D. L. Woodruff, and W. J. Hopp, "CONWIP: A pull alternative to kanban," *International Journal of Production Research*, vol. 28, no. 5, pp. 879–894, 1990.

[14] I. Duenyas and J. Wallace, "CONWIP assembly with deterministic processing and random outages," *IIE Transactions*, vol. 24, no. 4, pp. 97–109, 1992.

[15] W. J. Hopp and M. L. Roof, "Setting WIP levels with statistical throughput control (STC) in CONWIP production lines," *International Journal of Production Research*, vol. 36, no. 4, pp. 867–882, 1998.

[16] J. M. Framinan, P. L. Gonzalez, and R. Ruiz-Usano, "The CONWIP production control system: Review and research issues," *Production Planning & Control*, vol. 14, no. 3, pp. 255–265, 2003.

[17] S. Gstettner and H. Kuhn, "Analysis of production control systems kanban and CONWIP," *International Journal of Production Research*, vol. 34, no. 11, pp. 3253–3273, 1996.

[18] Y. Khojasteh-Ghamari, "A performance comparison between kanban and CONWIP controlled assembly systems," *Journal of Intelligent Manufacturing*, vol. 20, no. 6, pp. 751–760, 2009.

[19] R. Courant and F. John, *Introduction to calculus and analysis*. Springer, 1999, vol. 1.

[20] J. D. C. Little, "A proof for the queuing formula: $L = \lambda W$," *Operations Research*, vol. 9, no. 3, pp. 383–387, 1961.

[21] C.-B. Yan, M. Hassoun, and S. M. Meerkov, "Equilibria, stability, and transients in re-entrant lines under FBFS and LBFS dispatch and constant release," *IEEE Transactions on Semiconductor Manufacturing*, vol. 25, no. 2, pp. 211–229, 2012.